LASSO based Resample Model Averaging for Genetic Association Studies

Jeremy A. Sabourin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2013

Approved by:

Andrew B. Nobel

William Valdar

Yufeng Liu

J.S. Marron

Michael Wu

Abstract

JEREMY A. SABOURIN: LASSO based Resample Model Averaging for Genetic Association Studies.
(Under the direction of Andrew B. Nobel and William Valdar.)

Significance testing one SNP at a time has proven useful for identifying genomic regions that harbor variants affecting human disease. In theory, simultaneous modeling of multiple loci should help when considering complex diseases affected by multiple predictors. However, they are typically applied in an ad hoc fashion: conditioning on the top SNPs, with limited exploration of the model space and no assessment of how sensitive model choice was to sampling variability. Formal alternatives exist but are seldom used. When considering complex traits in humans, the genetic model is most often assumed to be additive only SNP effects. When non-additive effects such as dominance or overdominance are present, additive only models can be underpowered. We first present LLARRMA, a resample model averaging based method using the LASSO that allows for additive. It estimates for each SNP, the probability that it would be included in a multiple SNP model in alternative realizations of the data. We show that under simulations based on real GWAS data, that LLARRMA identifies a set of candidates that is enriched for causal loci relative to single locus analysis.

We next generalize the resample model averaging framework and present LLARRMA-dawg, a generalized resample model averaging based method using the group LASSO that allows for additive and non-additive SNP effects. We show that under simulations based on real GWAS data, that LLARRMA-dawg identifies a set of candidates that

is enriched for causal loci relative to LLARRMA in the presence of non-additive effects. We examine how the framework for LLARRMA-dawg can be extended to other problems where multiple model predictors are required to model the effects of a single variable.

The final portion of this dissertation describes additional information that one may explore from resample model averaging. Specifically, we examine how one can identify response specific variable relationships based on the models selected under resampling. This give the researcher further information about the predictors than the standard pairwise correlation structure which does not account for the response.

Acknowledgments

Much of the work described in this dissertation is collaborative, and I am very grateful for all of the help I have received. In particular, I would like to thank:

- My advisors Andrew Nobel and William Valdar, for their patience, dedication, and many insightful comments.
- My committee members Yufeng Liu, Steve Marron and Michael Wu, for their helpful criticism and suggestions.
- Ethan Lange and Leslie Lange for providing data and patiently explaining scientific questions and concepts.
- Andrey Shabalin, Jeff Roach and many others for their additional helpful comments.
- My friends and family for their support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This Dissertation consist of seven chapters. Each chapter addresses a problem or application in statistical genetics. Each is related to the general problem of how to identify genetic variants affecting a given disease trait. This is also commonly known as a "genetic associations" study.

This chapter provides the necessary genetics background (methods and terminology) to understand the research problems discussed in the later chapters. The chapter is laid out as follows. Section 1.1 describes the basic genetics topics that are important to genetic association studies. Section 1.2 describes the statistical literature on the human related research problems discussed. Section 1.3 describes the terminology and literature related to the use of model organisms. Section 1.4 describes how genetic information is commonly incorporated into a statistical model of association with a disease trait. Section 1.5 gives an overview of how receiver operator characteristic (ROC) curves are used to compare methods. The chapter ends with an overview of the research questions addressed in the dissertation (Section 1.6).

1.1 Basic genetic background

In order to understand fully the statistical modeling of genetic diseases, one needs a basic understanding of the genetics behind the model. We first discuss the standard

predictors for a genetic association study, followed by other genetic complications that often arise within genetic association studies.

1.1.1 DNA structure and SNPs

In the simplest models of association, a single outcome is predicted by a variable representing the genetic state of the individual. A common scenario models the state of individuals differing at a particular variant in their DNA (deoxyribonucleic acid). DNA carries the genetic instructions used for the development and functioning of all known living organisms (with the exception of RNA viruses). The information in DNA is stored as code consisting of four chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases are often referred to as nucleotides, and appear in pairs. Figure 1.1 displays the structure of the DNA molecule. The pairs come in one of two forms, A with T or C with G, and form units called base pairs. The DNA molecule is a long chain of these base pairs separated into long structures called chromosomes.

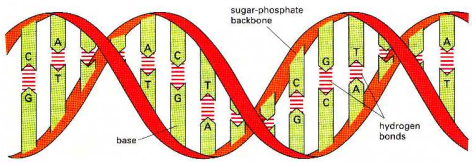

Figure 1.1: Displays the structure of DNA from (Ansari, 2001).

Human DNA comprises of about 3 billion base pairs split over 23 chromosomes, one of which is our sex chromosome. Figure 1.2 displays the chromosomes of a human male. More than 99 percent of these bases are the same in all humans. The locations that differ among individuals are the source of heritable variation in humans (eg. height, eye color, etc.). Each such location, generically referred to as a locus (plural loci), potentially harbors variation that affects a trait or disease (often called a phenotype), and is a candidate predictor for genetic associations. Along with many other organisms, humans are diploid, meaning that we contain two versions of each chromosome (known

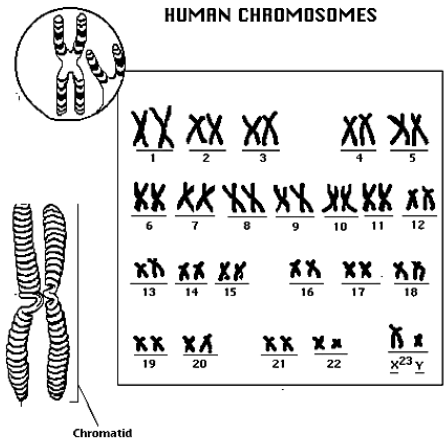

Figure 1.2: Representation of the 23 paired chromosomes of the human male; modified from (Access Excellence, 2009).

	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7
Chromatid 1	A	T	C	A	A	G	T
Chromatid 2	T	T	T	G	A	A	T

Table 1.1: The base pairs of an individual at 7 loci; we observe that 4 of the 7 loci we have different alleles on each chromatid and the remaining 3 have the same allele.

as homologues chromatids), a paternal copy passed down from our father and a maternal copy from our mother. As all humans contain two versions of their DNA, each locus contains two chemical bases (possibly the same). Table 1.1 displays the base pairs of an individual at 7 loci. Observe that 4 of the 7 loci have different alleles on each chromatid and the remaining 3 have the same allele. Each chemical base that is observed at a given location is referred to as an allele. The alleles at a given locus are the predictors we will consider for genetic association with a phenotype.

The information encoded in one chromatid is called a haplotype. It is often convenient to consider both haplotypes at once. The combination of the two haplotypes at a single locus is referred to as a genotype. It is most common that a locus where two individuals' DNA differs will has only two possible alleles, or nucleotides. Such loci are referred to as a single-nucleotide polymorphism (SNP). For now, we will only consider

loci that are SNPs. [We will investigate loci that are not SNPs in Chapter 4.] As a SNP contains only two possible alleles, it is common practice to call the less common allele in the population the "minor allele", Q, and the more common allele, q, the "major allele".

Given that we know the allele present on each haplotype at a given locus, we can determine the possible genotypes. The possible genotypes that can be observed are homozygous minor, QQ, heterozygous, Qq or qQ, and homozygous major, qq. At present, most techniques for identifying alleles are not capable of resolving haplotypes of an individual, but merely return the two alleles present at each loci. Thus, it is common practice to not distinguish between the heterozygous genotypes, Qq and qQ. With this simplification, the standard additive effect SNP predictor is defined as the count of the minor allele at a given locus, i.e., we code the unordered genotypes {qq, qQ, QQ} as {0, 1, 2}. With this representation of a SNP, one can begin to see how we may look to statistically model a phenotype as a function of the SNP to detect if the SNP has a significant additive effect on the phenotype. We consider more general predictors in Section 1.4.

1.1.2 Genetic models for phenotypic effects

With an understanding of the loci considered for causal variants, one must ask how these loci affect a phenotype. The underlying genetic models have been disputed, and exist in many different varieties. The most detailed model would explain phenotypic variations by including genetic variants, environmental effects, and potential interactions between and within these effects. Models with this level of detail are usually impractical to estimate and are rarely considered. Simpler models are easier to understand, collect data for, and model statistically. Below describes some of the historically relevant models that have been used, leading up to the model commonly used in today's literature.

Single variant mutations - Mendelian diseases

The origins of genetic association studies come from family based studies of Mendelian diseases, or diseases resulting from a single mutation in the structure of DNA, which cause a single basic defect with pathologic consequences. There are thousands of genetic diseases caused by a single mutation, but discovering which locus is the culprit is still a daunting task. As Mendelian diseases result from a single mutation, studying the DNA for multiple generations of a family infected by a disease allows for the identification of loci that potentially directly results in the disease. While family based studies have been successful for Mendelian disorders, there remained many diseases for which family designs were unsuccessful in identifying the underling genetic variants. These diseases became known as common diseases and their analysis erupted with the introduction of the genome wide association study (GWAS; WTCCC, 2007).

Simple single locus model - GWAS for common diseases

The simple single locus GWAS model assumes that there is a single SNP variant that has a causal effect on the phenotype. Unlike the Mendelian model, it is not assumed that the presence of a single allele determines the disease status of the individual, but rather it is assumed that the effect of the allele's count is additive, i.e., the effect of having two copies of the minor allele is double of that of having a single copy of it. The effect of locus x_i on phenotype y may be modeled as

$$y_i = \mu + x_{i,j}\beta_j + \epsilon_i$$

where μ is the phenotypic mean, β_j is the effect of locus x_j on y, and ϵ_i is a normal error. This simplistic model was one that was assumed in many analyses, and is consistent with many common diseases. This model is not often considered in today's literature, as most studies are examining complex traits, or traits that are effected by multiple loci

and potentially interactions of loci and environmental effects. Studies have shown that phenotypes for complex diseases require multiple variants to explain the differences of the phenotype (Su, Marchini and Donnelly, 2011). Even though it has become common practice to assume a more complicated model for complex diseases, it is still common that studies report only a single locus in their findings.

Simple multiple locus model - complex trait models

As studies have shown that the mendelian model is not valid for complex traits, many now assume a similar model with contains multiple loci. This model has multiple causal loci, but for mathematical convenience each locus acts with an independent additive effect. This may be represented by the the model

$$y_i = \mu + \sum_{j \in \mathcal{J}} x_{i,j}\beta_j + \epsilon_i$$

where $\mathcal{J}$ is the set of SNPs with true signals. This model has allowed researchers to account for a much larger amount of variability in complex phenotypes than that accounted for by the Mendelian model. Even though we assume a multiple locus model, it is still common practice to test loci one at a time. When considering the simple multiple locus model, we assume that there exists only a few loci that are truly causal. This means in statistical terms that we are assuming the model to be sparse.

Genetic dominance in the model

The simple models that we have considered thus far assumed that the SNPs have an additive effect on the phenotype. Unfortunately, not all genetic effects follow additive models. By only modeling additive effects, potential causal variants may be missed. To extend the model further, we remove the assumption of an additive effect to allow for dominant effects. A dominant effect, in its simplest form, means that having one copy of the minor allele has the same effect on the phenotype as that of two copies.

Under this definition, one may also consider recessive effects, meaning having one copy of the minor allele having the same effect as not having a copy (one may think of a recessive effect as the major allele is dominant). We will consider a more general notion of dominance, where by dominance we mean any deviation from an additive model.

The general notion of dominance can be harder to detect, but allows for models that more closely resembles observed effects. The common genetic dominance model for an individual SNP is to assign two values a and d for the additive and dominant effect aspects respectively. We assume that the homogeneous minor allele state, QQ, will have an effect with a value of a, and the homogeneous major allele state, qq, will have an effect of $-a$, and the heterogeneous state, Qq or qQ, will have an effect of d. Under this setting, the coding of the genotypes would differ from the standard 0, 1, 2 coding but be coded as -1, 0, and 1 for qq, qQ, and QQ respectively. This can be represented in a single locus regression model for locus j as

$$y_i = \mu + a_j x_{i,j} + d_j I(x_{i,j} = 0) + \epsilon_i.$$

Table 1.2 displays how one would write out several models under this notation. This notation generalizes the additive model to allow a dominant model; it is easily observable that when there is no dominant effect, ie. $d = 0$, that it simplifies to the standard additive effects centered at the heterogeneous state. It is also observable that if $d = a$ ($d = -a$) that this modeling simplifies to the simplistic definition of a dominant (recessive) effect. To illustrate how different dominant effects can be from the standard additive effects, Figure 1.3 displays the non-additive dominance type effects which the model could represent for different values of a and d.

	Additive	Dominant	Recessive	General Dominance
d	0	a	-a	$a \times \alpha \in \mathbb{R}$
qq	-a	-a	-a	-a
qQ	0	a	-a	αa
QQ	a	a	a	a

Table 1.2: Displays how one models different levels of dominance under the a, d notation.

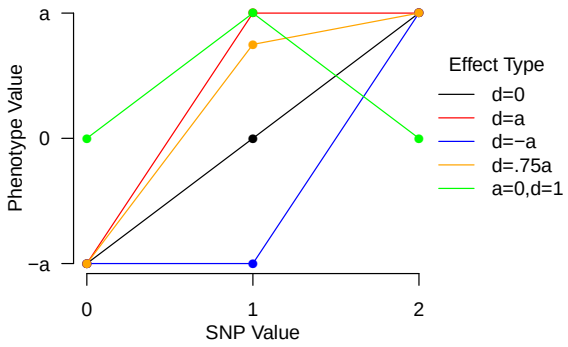

Figure 1.3: Comparison of dominant effects to the additive model.

More general models

As previously mentioned, the most detailed model considered would contain both genetic variants, environmental effects, and interactions between and within these effect. To account for as detailed of a model containing only genetic effects, one may extend the multiple locus model to include not only the dominance predictors, but also interactions. Models with interactions are sometimes considered, but are harder to model statistically; even assuming a sparse model with only considering pairwise interactions, a very large number of interactions are possible leading to a very large number of predictors in the model (Balding, 2006). Given the sparse set of causal variants we could fit a model to include interactions easily, but in this setting it would be more common to perform a haplotype analysis. A haplotype analysis examines the unique combinations of each loci, as a haplotype, for association with the phenotype rather than individual effects and the interactions between them.

	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7
Chromatid 1	A	T	C	A	A	G	T
Chromatid 2	T	T	T	G	A	A	T
Genotype	AT	TT	CT	AG	AA	AG	TT
Phased Genotype	A/T	T/T	C/T	A/G	A/A	G/A	T/T

Table 1.3: The base pairs of an individual at 7 loci. The lower half of the table displays the genotypes and phased genotypes of the individual at each loci. The allele sequence of each chromatid are examples of haplotypes. While the genotype row give no indication of the underling haplotypes, in the phased genotypes we notice that the first allele in the phased genotype always corresponds to chromatid 1. This means that the sequence given by the first allele in the phased genotypes forms the haplotype from chromatid 1.

When the contributing loci in a genetic model have been identified, rather than examine the interactions of the loci, it is more common to phase the data and examine the individual haplotypes created by the set of loci. By phasing the data, we mean to analyze the haplotypes in the population and use this information to reconstruct haplotypes from the genotype information. Table 1.3 revisits the data from Table 1.1 and displays the standard genotype information along with the phased genotypes obtained after reconstructing the individual haplotypes. As none of the research presented here is directly related to the use of haplotype phasing, no more details are given here (see Browning (2008) for further details). With the haplotypes reconstructed, a haplotype model would assume that each observed haplotype has its own effect. This leads to a model which can easily be modeled statistically, assuming that you can identify the proper loci to include in a haplotype. When selecting the loci to include in the haplotype model, one needs to consider how fast the size of the model grows, i.e., including p SNPs results in 2^p possible haplotypes in the model. For this reason, haplotype models are most often considered for small regions where it is suspected that there are multiple causal loci, which may not have been validated. This model may include spurious loci and may also be missing important loci.

1.1.3 Linkage disequilibrium

One of the largest hindrances to a genetic association study is the pattern of correlation, or linkage disequilibrium (LD), between SNPs. LD is considered a measurement of how likely combinations of alleles would be under the assumption of random formation of haplotypes. A common measure of LD is the square of the correlation (Balding, 2006). SNPs that are close together are often in high LD, and this causes a confounding relationship between the SNPs close to a causal SNP and the phenotype. This often results in a set of SNPs that form a "cloud" of statistically significant SNPs when modeled based on single locus methods, as displayed in Figure 1.4. Many standard statistical methods are unable to distinguish which SNP is the causal SNP within a set of SNPs in high LD as they fail to incorporate LD information into the analysis (Balding, 2006).

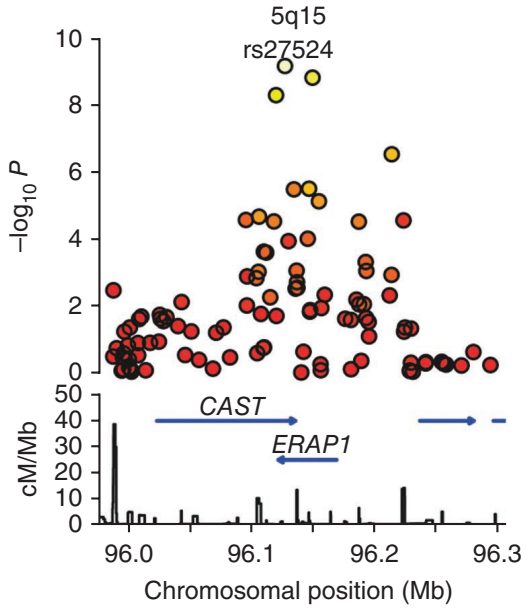

Figure 1.4: An example of the resulting $-\log_{10}$(p-values) from a single locus regression in a region of high LD from Strange et al. (2010)

.

However LD can also aid genetic analyses, particularly for the imputation of missing genetic data. But before we discuss how LD can be used here, we will first review the causes of LD.

LD is the occurrence of some combinations of alleles in a population more or less often than one would expect from a random formation of haplotypes from alleles based on their frequencies under Hardy-Weinburg (HW) equilibrium. HW equilibrium states that both allele and genotype frequencies in a population remain constant, that is, they are in equilibrium, from generation to generation unless specific disturbing influences are introduced. In the simplest case of HW equilibrium, we observe a single locus with two alleles. The minor allele is denoted A and the major allele a; their frequencies are denoted by p and q respectively. If the population is in equilibrium, then the homozygote genotypes AA and aa are observed with frequency p^2 and q^2 respectively in the population, and the unphased heterozygote genotype aA is observed with frequency $2pq$.

The combinations of alleles that result in LD can most easily be explained by the process of how our DNA is passed down from our parents. Each copy of a chromosome was obtained from one of our parents, but this copy is not an exact copy of one of their chromosome. Rather, it is a combination of their chromosomes. When producing sex cells, the process of meiosis creates a new copy of each chromosome that is a mixture of their two copies. The new chromosomes are not highly mixed, but rather a crossover event occurs a few times per chromatid. In 1919, J. B. S. Haldane proposed a statistical model for genetic recombination, the stochastic process that occurs during cell meiosis whereby the paternal and maternal chromatids (the chromatids obtained from their mother and father) of each chromosome exchange some of their genetic material to form recombinant chromatids; the resulting set of chromatids being used to form two sex cells as displayed in Figure 1.5. According to the Haldane model, crossovers

(the point of genetic exchange; also known as recombination breakpoints) occur along the chromosome as a Poisson process. In the simplest theoretical representation, the expected number of crossovers for a chromosome is 1, that is, one point of exchange between the maternal and paternal chromatids. In fact, Haldane realized that some chromosomes undergo more crossovers than others. He therefore defined the unit of *genetic distance*, the *Morgan*, as the expected number of crossovers between two loci. Specifically, if genetic distance is measured in Morgans (M), then the rate of the Poisson process is 1 for a 1M chromosome, and 1.5 for a 1.5M chromosome.

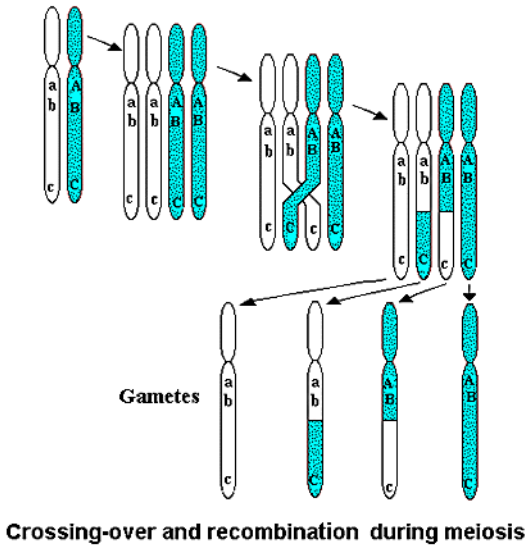

Figure 1.5: An illustration of a crossover adapted from Morgan et al. (1915).

As crossover events are not numerous in meiosis, we pass along long sequences of DNA that are the same as our previous generation which results in some correlation or LD. As a population ages, the level of LD slowly weakens due to the random process of crossovers. As the human population is still relatively young, our DNA exhibits strong LD in areas that crossover events are less common.

Now to get back to how we can use LD. With an understanding of the crossover rates from studying LD and reference population haplotypes gained from the HapMap

(Tanaka, 2009) or 1000genomes (Siva, 2008) projects, one can phase an individuals haplotype. This is done by using the neighboring loci to infer the probability of the haplotype that is present at a given location based on comparison to the reference haplotypes and using the knowledge of recombination rates from LD to estimate the probability of a recombination at this location. The ability to model this information through the use of Hidden Markov models provides an accurate method to estimate phased haplotypes (Scheet and Stephens, 2006). These phasing methods also provide the basis for estimating the information needed to impute, or fill in, missing values in genotype or haplotype data.

Missing data and imputation

SNP data within a GWAS will most always include combinations of loci and individuals where the genotype is unknown or uncertain. To avoid a potentially wasteful complete cases analysis, it is common to impute the missing genotypes using a program such as MACH (Li et al., 2010), IMPUTE (Howie, Donnelly and Marchini, 2009) or fastPHASE (Scheet and Stephens, 2006), and analyze the partly-imputed data as if it were fully observed. Imputation methods are typically based on reconstruction and phasing of inferred haplotypes. Let us consider the SNP matrix $\mathbf{X}$ which may be divided into its missing and observed elements as $\mathbf{X} = \{\mathcal{X}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$.

In this setting, imputation methods such as fastPHASE (Scheet and Stephens, 2006) model the joint distribution of the missing genotypes $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ includes additional information used in the imputation (e.g., priors or additional haplotype data from HapMap (Tanaka, 2009) or 1000Genomes (Siva, 2008)). However, most GWAS studies do not use this joint distribution directly. In most GWAS, they replace $\mathcal{X}_{\text{mis}}$ with an element by element point estimate of $\widehat{\mathcal{X}}_{\text{mis}}$ that is constructed based on the marginal conditional distributions of individual missing elements. The standard imputation methods for SNP data replace $\mathcal{X}_{\text{mis}}$ by either the "dosage" effect, $\widehat{\mathcal{X}}_{\text{mis}}^{\text{dose}}$, whose

elements are defined as the expectation of the minor allele count $\hat{x}_{ij} = \mathrm{E}(x_{ij}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$; or by a "hard" imputation, $\widehat{\mathcal{X}}^{\text{hard}}_{\text{mis}}$, whose elements are imputed as the genotype with the maximum posteriori probability

$$\hat{x}_{ij} = \underset{g\in\{0,1,2\}}{\operatorname{argmax}}\, p(x_{ij} = g|\mathcal{X}^{\text{obs}}, \boldsymbol{\omega}).$$

The simplest approach to modeling missing genotypes within GWAS is to estimate $\mathcal{X}_{\text{mis}}$ as either $\widehat{\mathcal{X}}^{\text{dose}}_{\text{mis}}$ or $\widehat{\mathcal{X}}^{\text{hard}}_{\text{mis}}$ and then assume that $\widehat{\mathbf{X}} = \{\widehat{\mathcal{X}}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$ was complete. This plug-in approach underestimates variability as it fails to incorporate uncertainty about the imputation (Little and Rubin, 2002). Zheng et al. (2011) show that when modeling effects at a single locus, using these plug-in approaches reduces the power to detect causal SNPs a negligible amount when the imputation accuracy is reasonably high. Nonetheless, ignoring imputation uncertainty could be more problematic in multiple locus settings. An example of this is if the joint posterior distribution of haplotypes $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ differs substantially from joint distribution implied by the product of marginal posteriors $\prod_{ij\in\mathcal{X}_{\text{mis}}} p(x_{ij}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ (eg, Servin and Stephens, 2007). A natural way to incorporate imputation uncertainty is through multiple imputation (Little and Rubin, 2002).

By multiple imputation we mean a method that calculates the joint posterior distribution of the data and randomly draws the values of the missing data from this distribution to create a version of the imputed data. This random draw would be repeated K times to create K different version of the imputed data, rather than simply plugging in a single value like the plug-in imputation methods had. When using multiple imputation to draw inference about a parameter θ, it is most common that the desired statistic, $\hat{\theta}$, would be computed for each imputed version of the data providing $\hat{\boldsymbol{\theta}} = (\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(K)})$. The final inference on θ would based on an aggregate of $\hat{\boldsymbol{\theta}}$. Little

and Rubin (2002) show that the combined analyses of the different imputed versions of the data can be used to account for the uncertainty of the unobserved data.

1.2 Overview of standard statistical methods for Human GWAS

A genome wide association study (GWAS) has become the most common experiment for understanding the genetics of diseases. Recent GWASs have investigated hundreds of thousands of single nucleotide polymorphisms (SNPs) for associations with a phenotype. Proper statistical modeling of the data is important when identifying SNPs that are associated with a phenotype. When analyzing the GWAS data it is important to consider the underlying genetic model in order to obtain the best possible results. Regression modeling has become the most commonly used statistical tool used in GWAS studies. The specific regression technique that has been used is single locus regression models that fit a model for each SNP in the data set individually. The $-\log_{10} P$ values (hereafter referred to as "logP") of the individual regressions are used to identify loci of significant association with a phenotype. These single locus regression methods have allowed researchers to investigates genetic association studies with large numbers of loci in a single study (WTCCC, 2007).

The single locus regression approach to GWAS analyses have allowed researchers to analyze large amounts of data quickly. Unfortunately, performing these single locus based analyses on each SNP assumes that the SNPs are independent, but LD has shown that SNPs are often highly dependent. Ignoring the LD often has negative effects on the single locus regression models. In regions of high LD, they often find large quantities of SNPs to be statistically significant even though there are likely only one or a few SNPs in the region that have a true association with the phenotype. These large clouds of significant loci, referred to as "hit regions", are often hard to interpret, and in practice may lead to the end of statistical analyses on the data. While

there are many proposed methods for analyzing GWAS data, single locus regressions are essentially the only method used in practice, and many practitioners who use it do not like the overall performance in regions of localized LD. Researchers often ask for the development of methods that better handle the localized LD which hinders the single locus method. There is thus great value in developing principled approaches to discriminate true from false associations in hit regions. This is further emphasized by sure independence screening (SIS) (Fan and Lv, 2008), where it is theoretically shown that prescreening the data for the set of most correlated predictors, or hit regions, and performing more detailed analysis on the selected set of predictors still selects the true set of predictors with probability tending to one.

1.2.1 Statistical analyses for hit regions

Statistical analyses run after selecting hit regions of top SNPs are often of an *ad hoc* nature. They typically involve fitting further regressions that condition on "top" locus that appear to be most strongly associated. The conditioning is used in order to rule out correlated neighbors of the top locus or rule in suspicions of an independent second signal. In ad hoc conditioning, rarely are there formal considerations of the fact that the association of the top locus is often insignificantly different from that of its correlated neighbors, and that while its association with the phenotype is probably stable to sampling error, its superiority in association over its neighbors is probably not. This inherent instability of the relative strengths of association between confounding loci makes such strategies extremely high risk: a slightly different sampling of individuals could demote the conditioning locus, resulting in an alternative conditioning locus and potentially lead to drastically altered conclusions. This approach becomes more precarious still when some of the loci are themselves known with varying certainty, their genotypes having been partially or wholly imputed (Zheng et al., 2011), so that

the weakness of association is now also a function of imputation uncertainty unrelated to the phenotype (eg, Servin and Stephens, 2007).

Joint modeling of all loci through multiple regression seems like an attractive alternative to the single locus regressions because it accounts for the LD of the data (Balding, 2006). However, standard multiple regression is often unsuitable because even when the number of loci p is much less than the number of individuals n, LD creates multicollinearity that prevents meaningful estimation of locus effects. Stepwise multiple regression techniques can be used to fit such models, formalizing the ad hoc conditioning approach, and thus also inheriting its weaknesses. Further, a model selection procedure that selects a single set of active loci typically provides no indication of the sensitivity of the set of loci with respect to, for example, sampling variability and imputation. This makes the set of selected loci form stepwise procedures a statistic that is obscure at best and misleading at worst. Bayesian approaches offer a coherent perspective by formally accounting for uncertainty in the model choice, the estimation of effects, and imputation uncertainty (Stephens and Balding, 2009). However, these are often computationally intensive, and require formal statements of prior belief relating to the number of causal variants and their effects that many analysts are unprepared or unwilling to specify.

Penalized regression models provide an alternative that does not require a commitment to Bayesian learning. Placing a penalty on the size of coefficients in the multiple regression objective leads to moderated estimates of coefficient effects. This allows for stable estimation when many predictors are in the model. In particular, the LASSO (Tibshirani, 1996) penalizes the absolute value of each coefficient subject to a penalty parameter λ, resulting in some effects being shrunk to exactly zero. This results in a "sparse" model in which only a subset of effects are active. Increasing the level of penalization leads to greater sparsity, effectively making λ a continuous model selection

parameter. Recent computational advances in the fitting of LASSO-type models have made them more practical for analysis of large scale genetic data (eg, Wu et al., 2009). Nonetheless, as a tool for modeling effects at multiple loci, the LASSO leaves important questions unanswered. One problem is how to select λ. This is typically approached by criteria-based methods (Zhou et al., 2010; Wu et al., 2009), such as AIC and BIC, empirical measures of predictive accuracy (such as cross validation; Friedman, Hastie and Tibshirani, 2010), and criteria aiming to control type I error (such as permutation; Ayers and Cordell, 2010). Another issue is how to characterize uncertainty in the model choice given parameter λ. Although LASSO moderates estimated effects through shrinkage, it is no better than stepwise methods in that it ultimately selects a single model (or single "path" of models, when λ is varied), and thus states with absolute confidence a statistic that could in fact be highly sensitive to sampling variability.

One intuitive way to characterize variability of model choice is to estimate a model inclusion probability (MIP) for each locus. Whereas a Bayesian approach would formulate this as a posterior probability that conditions on both the observed data and prior uncertainty in model choice, a frequentist alternative is to formulate the MIP as the probability a locus would be included in a sparse model under an alternative realization of the data. Valdar et al. (2009) proposed an approach that applied forward selection of genetic loci to resamples of the data and defined the resample MIP (RMIP) as the proportion of resampled datasets for which a locus was selected. This resample model averaging (RMA) approach used either bootstrapping (ie, "bagging") or subsampling (ie, "subagging"), and followed an earlier application to genome wide association in Valdar et al. (2006) and work on general aggregation methods by Breiman (1996) and Bühlmann and Yu (2002). Independently, Meinshausen and Bühlmann (2010) proposed "stability selection", which powerfully combines subagging with LASSO shrinkage to produce a set of frequentist MIPs at each specified λ.

1.2.2 Stability selection

Stability selection (Meinshausen and Bühlmann, 2010) combines the use of subsampling and LASSO penalized regression in order to account for sampling uncertainty in the problem of model selection. The use of stability selection has been widespread do to its generality, and has proved to be useful in many areas.

The idea behind stability selection is not to apply the LASSO to a data set and take the selected model (or path of models) to be the true model. Rather, stability selections seeks to apply the LASSO to many subsamples (of half of the data) and select the variables which are selected within a large percent of the data over a user determined set of penalty parameters. The variables which have a high MIP (above some threshold) are considered stable variables and should be included in your final set of variables. Under some reasonable assumptions (see further discussion in the next subsection), Meinshausen and Bühlmann (2010) establish a bound on the expected number of noise variables with MIPs above a threshold (between .5 and 1) based on the expected number of variables included on each subsample. This bound can be useful in helping the user decide on the set of penalty parameters to be used and/or the threshold for stable variables.

Recently, stability selection has been revisited by Shah and Samworth (2011), who restate stability selection in a more general framework. They also made minor modifications to the stability selection procedure. The changes to the procedure along with some distributional observations on the inclusions probabilities have resulted an analogous version of the bound on the number of falsely selected variables for their setting.

Shah and Samworth (2011) modified the subsampling scheme of stability selection to include $K/2$ complementary pair subsamples rather than K subsamples as used by stability selection, i.e., they use subsamples of size $\lfloor n/2 \rfloor$, with indices $(\mathcal{N}^{(2j-1)}, \mathcal{N}^{(2j)}; j =$

$1, \ldots, K/2$) where $\mathcal{N}^{(2j-1)} \bigcap \mathcal{N}^{(2j)} = \emptyset$. They refer to the modified method as complementary pairs stability selection (CPSS). This modification falls into the setting of the original setup, thus all theory from stability selection (Meinshausen and Bühlmann, 2010) still holds.

Shah and Samworth's motivation for the complement pairs comes from the the proofs on stability selection. The proofs directly use simultaneous selection probabilities for complementary pairs, even though stability selection does not sample the complementary pair. Another large distinction from Meinshausen and Bühlmann (2010) is that rather than use false (or noise) variables Shah and Samworth consider "low selection probability" variables, where the set of low probability selection is defined as

$$L_\theta = \{k : p_{k,\lfloor n/2 \rfloor} \leq \theta\}$$

where $p_{k,n}$ is the selection probability of variable k under the considered variable selection procedure based on n individuals. By considering L_θ rather than the set of noise variables, Shah and Samworth are able to reduce the number of assumptions for their theorem for the number of falsely selected low selection probability variables. Shah and Samworth argue that one can consider a noise variable as a low selection probability variable to obtain a bound on false discoveries. While this may be appropriate in some settings, it may not hold in the setting of a GWAS hit region. This is because a SNP that is in high LD with the causal SNP may have a selection probability higher than the threshold used to define a low selection; while you can raise the threshold to address this, a causal SNP with a low MAF may then fall into the set of low selection probability variables even though it is not a noise variable.

Detailed overview of stability selection

To best understand the procedure of stability selection, let us consider data $\mathcal{D} = \{\mathbf{y}, \mathbf{X}\}$ with $\mathbf{y} = \{y_1, \ldots, y_n; y_i \in \mathbb{R}\}$ being a function of $\mathbf{X} = [X_i, \ldots, X_n], X_i \in \mathbb{R}^p$ and indices $\mathcal{N} = \{1, \ldots, n\}$. We assume that the relationship between $\mathbf{X}$ and $\mathbf{y}$ can be modeled by a regression model, such as a linear model

$$\mathbf{y} = \mu + \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where μ is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the effects of the p variables of (X), and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$. Assume that of the p variables in $\mathbf{X}$, only a subset are 'true' predictors, i.e., non-zero coefficients. If we define the set $S = \{k : \beta_k \neq 0\}$ as the set of true predictors, out goal is to use $\mathcal{D}$ to select the set S. Meinshausen and Bühlmann define a selection probability as follows:

Let $\mathcal{D}^{(k)}$ be a random subsample, drawn with replacement, of $\mathcal{D}$ of size $|\mathcal{N}^{(k)}| = \lfloor n/2 \rfloor$, such that $\mathcal{N}^{(k)} \subset \mathcal{N}, k = 1, \ldots, K$. For every set J, the probability of being selected in the set $\hat{S}(\mathcal{D}^{(k)})$ is given by

$$\hat{\Pi}_J = P^\star\{J \subset \hat{S}(\mathcal{D}^{(k)})\}$$

where the probability $P^\star$ is with respect to random subsampling.

Model selection for stability selection is performed on each subsample k by the use of the LASSO penalty (Tibshirani, 1996) which gives estimates

$$\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)}) + \lambda \sum_{j=1}^{m} |\beta_j| \right\}, \tag{1.1}$$

where $\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)})$ is the log-likelihood of $\boldsymbol{\beta}$ for data $\mathcal{D}^{(k)}$, and λ is a penalty parameter. From these estimates we can define the set of selected predictors $\hat{S}_k^\lambda = \hat{S}^\lambda(\mathcal{D}^{(k)}) = \{k :$

$\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) \neq 0\}$. Aggregating over K subsamples, we obtain the estimate of $\hat{\pi}_J$ as

$$\hat{\pi}_J^{\lambda} = \frac{1}{K} \sum_{k=1}^{K} I(J \in \hat{S}^{\lambda}(\mathcal{D}^{(k)})).$$

To select a final set of variables to include, Meinshausen and Bühlmann select the stable variables for a cut-off π_{thr} with $0 < \pi_{\text{thr}} < 1$ and a set of regularization parameters Λ. The set of stable variables is defined as

$$\hat{S}^{\text{stable}} = \{j : \max_{\lambda \in \Lambda}(\hat{\pi}_j^{\lambda}) \geq \pi_{\text{thr}}\},$$

that is we include the set of variables which were included in a proportion of the subsample models higher then π_{thr}. Note that in practice Λ is most often chosen to be a single value of λ, but the selection of this value is a difficult problem.

Meinshausen and Bühlmann provide theoretical properties for stability selection. These include a bound for the number of falsely selected stable variables and results on consistent variable selection. Theorem 1.1 states the bound for the number of falsely selected null variables. Define $q_{\Lambda} = E(|\cup_{\lambda \in \Lambda} \hat{S}^{\lambda}|)$, i.e., the expected number of variables selected over $\lambda \in \Lambda$.

Theorem 1.1. *Theorem 1 (Meinshausen and Bühlmann, 2010) Let q_{Λ} be as above, and define $N = \{k : \beta_k \neq 0\}$. Let $V = |N \cap \hat{S}^{\text{stable}}|$ be the number of falsely selected variables with stability selection.*

Assume that the joint distribution the null variables, $\{I(k \in \hat{S}^{\lambda}), k \in N\}$ is exchangeable for all $\lambda \in \Lambda$. Also assume that the original selection procedure is no worse than random guessing for all $\lambda \in \Lambda$. Then, the expected number of falsely selected variables for $\pi_{\text{thr}} \in (0.5, 1)$is bounded by

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_{\Lambda}^2}{p}.$$

Theorem 1.1 can be useful in practice for the often difficult choice of the best value of λ, or $\lambda \in \Lambda$ if considering a range of values. Unfortunately, when it comes to the use of genetic data, the exchangeability assumption will be violated by LD. This makes the use of this bound for selecting λ when applying stability selection to a hit region potentially ambiguous.

Meinshausen and Bühlmann continue their theoretical work by showing that under some assumptions that stability selection has consistent variable selection, ie. $P(\hat{S} = S) \to 1$ as $n \to \infty$. The assumptions of the consistent variable selection are rather strong. In order to obtain consistent variable selection in a less restrictive setting, Meinshausen and Bühlmann (2010) introduce the randomized LASSO.

The randomized LASSO (Meinshausen and Bühlmann, 2010) is a new generalization of the LASSO. Rather than penalizing each variable by its coefficients absolute value, $|\beta|$, by a weight proportional to λ, the randomized LASSO changes the penalty λ to a randomly chosen value in $[\lambda, \lambda/\alpha]$ where α is a weakness parameter with $\alpha \in (0, 1]$. The randomized LASSO estimator $\hat{\beta}(\lambda, \alpha; \mathcal{D})$ is

$$\hat{\beta}(\lambda, \alpha; \mathcal{D}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\beta}; \mathcal{D}) + \lambda \sum_{j=1}^{m} \frac{|\beta_j|}{W_j} \right\},$$

where $W_j \sim U(\alpha, 1)$ is the weighting parameter with $\alpha \in [0, 1]$. While we have defined the randomized LASSO estimator for the entire data set, it is most appropriately used for multiple resamples of the data. This is because it would not make sense to apply the penalty a single time to the full data set as W randomly down-weights some predictors relative to others.

Since the random penalties are regenerated for each subsample, this produces a randomized re-weighting that can help deal with the shortcoming of the LASSO when you have a set of highly correlated predictors. When a group of variables are highly

correlated, the LASSO tends to select just a single variable, or favors the one variable, and ignores the others by setting them equal to zero. Thus, the use of the randomized LASSO allows for the identity of favored predictors to shift within correlated groups between subsamples, thus counteracting the favoritism of the LASSO. Meinshausen and Bühlmann (2010) advocate choosing $\alpha \in [0.2, 0.8]$ with lower values producing a more sparse set of stable variables.

Stability selection in genetics

Recently, stability selection has been applied to genetic data. Alexander and Lange (2011) implement a standard version of stability selection for genome-wide associations studies. They found that stability selection lacks power to detect true associations when compared to the standard single locus regression models on a standard set of GWAS data from WTCCC (2007) and simulated data sets. While stability selection was underpowered, there are many aspects of the implementation of stability selection that makes one suspect that the use of stability selection for genetic association is 'sold short' by the results of Alexander and Lange (2011). Before going into details of why one may feel this way, lets examine what was done in the paper.

Alexander and Lange (2011) produce two versions of stability selection, the first was a rather standard implementation of stability selection that uses Theorem 1.1 to select the value of λ. While Alexander and Lange do address the fact that LD will violate the exchangeability condition of the theorem, they mention that Meinshausen and Bühlmann (2010) speculate that the Theorem 1.1 may hold under more general conditions. Although they do perform a permutation test to validate that the false discovery bound obtained from SS theorem 1 is reasonable, they performed only one permutation per data set. Alexander and Lange (2011) also use the Theorem 1.1 in a non standard manor due to their misconseption about the algorithm used to fit the LASSO. They first select the π_{thr} value and desired number of false positives and then

use Theorem 1.1 to find the average number of variables that should be included, q. For the first 10 subsamples, they find the λ that includes q predictors. The lambda that is then used for stability selection is the average of the λs from the first 10 subsamples.

The setting which Alexander and Lange tested stability selection was not a setting that would motivate the use of a multiple locus method. Specifically, the simulations run by Alexander and Lange should favor single locus methods as the true signals are simulated on separate chromosomes, essentially making each of them single independent signals. With the relatively large sample size and very large effect sizes that are used in the simulations, single locus methods would perform well unless the true variables were to be in an area of extremely high LD. Unfortunately, we do not have any information about the LD within the independent regions where the true SNPs were located, and thus are unable to accurately evaluate if stability selection is even being compared to single locus regression in a fair setting.

While the standard implementation of stability selection in Alexander and Lange (2011) was found to lack power in GWAS analyses, He and Lin (2011) were more successful in incorporating stability selection for GWAS data. He and Lin (2011) propose to use stability selection for case/control data with the LASSO model selection procedure replaced by a three stage iterated sure independence screening (ISIS) (Fan and Lv, 2008) procedure, which they refer to as GWASelect. The first iteration is a marginal SIS procedure and the second and third iterations are conditional SIS procedures to ensure no variables may have been missed in the first screening.

GWASelect performs well in their simulation studies. But, one may argue that the simulation settings are too easy to detect true signals. Specifically, the SNPs in the simulations were chosen so that they have a minor allele frequency (MAF) of 0.3. As for a GWAS study a MAF of 0.05 or higher is considered a common allele, He and Lin have have required a very common allele in their simulation. Beyond the high MAF

of the true SNPs in the simulations, the effect sizes are also chosen to be much larger than one would expect to see in real data. The effect sizes for their three simulations were 0.35, 0.3, and 0.5 for their first, second, and third simulation sets respectively. To compare these effect sizes, consider that the effect sizes used in the research presented in this dissertation run between 0.13 and 0.25 with a mean of 0.22.

He and Lin also define a true positive as any SNP within 50SNPs of the true SNP with $r^2 > 0.05$, and a false negative cluster as any set of SNPs within 10SNPs of each other; each false positive cluster counts as only 1 false positive. This allows us to see that GWASelect does a good job of selecting regions close to the true SNP, but does not actually give any information about the methods ability to select the correct SNP.

While it may not be clear how well stability selection or modifications of it will truly perform on genetic hit regions from GWASs, it is clear that it is a legitimate competing method for multiple locus methods. Thus, stability selection is an important competing method for the multiple locus methods discussed in the later chapters.

1.3 Model Organisms and Association Mapping

Although the main focus of this dissertation is on human genetics, we have adapted the methods presented for use with model organisms. Model organisms are non-human species that are studied with the goal of better understanding biological phenomena, including biological mechanisms relevant to human disease (Palmer and de Wit, 2011). They are studied with the hope that the data and theories generated will be applicable to other organisms (Ankeny and Leonelli, 2011). The use of model organisms has played a vital roll in much of our understanding of heredity, development, and molecular processes that are required to understand phenotypic diversity in humans and other organisms (Müller and Grossniklaus, 2010). Much of the success of model organisms has resulted from there experimental advantages. Model organisms (in particular rodents,

flies, and worms) are relatively inexpensive to gather, transport, maintain and manipulate experimentally. They also have short generation times, high fertility rates, and high susceptibility for genetic manipulation (Ankeny and Leonelli, 2011). This leads to an ideal setting to breed and maintain in large numbers under laboratory conditions.

Many organisms have been used as model organisms over the years. Some of the most common include the fruitfly (drosophila), the nematode worm, yeast, the mouse, and the rat. We will focus on the use of the mouse, which has become the leading mammalian model organism (Müller and Grossniklaus, 2010). The mouse is closely related to humans (Waterston et al., 2002), and shares many developmental strategies and diseases. These include cancer, hypertension, diabetes, and glaucoma. Many other deceases which do not normally arise in mice can be induced by manipulation of genetics and/or environments. These include diseases such as Alzheimer's disease and cystic fibrosis.

In a recent perspective paper (Aitman et al., 2011), the authors addressed the question of, 'Why do we still need model organisms to understand human disease?'. Many important observations where made within the paper. Some of the relevant response are discussed below. Timothy J. Aitman addressed the fact that loci obtained from GWAS explain a relatively low proportion of the heritability for a disease or trait, making it difficult to establish the mechanism which controls the phenotype. With the use of model organisms, we have been able to identify the underlying gene for some phenotypes which may have not been accomplished without model organisms. One such example of this the identification of $Cd36$, which was identified as an insulin resistance gene in rats and then humans (Aitman et al., 1999). Gary A. Churchill discusses how the challenges posed by population stratification, rare alleles, uncontrolled environments, and constraints on experimental validation can be circumvented by the use of model

organisms. He also points out that Disease alleles do not need to be identified in humans to have relevance in human health.

To emphasize the importance of the answers of Timothy J. Aitman and Gary A. Churchill discussed above, we examine how mice have been used to directly effect human genetics. Although the same polymorphisms are not expected to exist in both mice and humans, the genes involved with many phenotypes are expected to exist in both populations (Palmer and de Wit, 2011). The assumptions of this gene relationship approach include that the genes have similar functions with respect to complex traits in mice and humans, that these effects are reasonably robust to genetic context (e.g. strain background in mice and genetic diversity in humans), and (only a small fraction of the genes in the genome have the ability to modulate the trait of interested (Palmer and de Wit, 2011). The final assumption is important because this discovery of genes using mouse studies will be most valuable if the total number of genes that influence a given trait is small, so that identifying such genes is significantly better than arbitrarily selecting a gene. An example gene in mice which has been related in humans is Casein kinase 1 epsilon (Csnk1e). Csnk1e was identified because it was differentially expressed in mice with high and low locomotor response to methamphetamine. Association between polymorphisms in CSNK1E and the subjectively rewarding effects of stimulant drugs in human volunteers in a laboratory-based study found a SNP in CSNK1E (rs135745) that was associated with the effects of amphetamine (Veenstra-VanderWeele et al., 2006), providing some support for the proposed relationship between the locomotor response to stimulant drugs. Another example is Hu et al. (2012), where the application of a network analysis on independent human/mouse gene-expression datasets resulted in reproducible sets of genes associated with metastatic disease in both mice and humans.

Another example of the human/mouse gene relationship is Mervis et al. (2012). They examined the role of one commonly duplicated or deleted gene in separation anxiety in chromosomal region 7q11.23. This region is known to cause neurodevelopmental disorders with contrasting anxiety phenotypes. Examining mice that had varying number of copiess of Gtf2i (a gene in this region) were examined and it was found that relative to mouse pups with one or two copies of Gtf2i, the pups with additional copies displayed a significant increase in maternal separation-induced anxiety. The study was able to link the copy number of a single gene from 7q11.23 to separation anxiety in both mice and humans, highlighting the utility of mouse models in dissecting specific gene functions for genomic disorders that span many genes (Mervis et al., 2012).

1.3.1 Historical Overview of Model Organisms in Biomedical Sciences

The most historical model organisms used began with the analysis of a single inbred line, where by inbred we mean that each individual in the line are genetically identical. These simple observational analyses allowed for better observation of phenotype values as you are able to average over many inbred individuals. To help understand the genetic effects for these diseases, recombinant chromosome substitution lines were proposed. They have had a long history for use in wheat breeding (Cavanagh et al., 2008). Typically chromosome substitution sets involve all chromosomes (except one) being derived from a recurrent parent and the remaining chromosome from a donor parent. Although effective, the resolution to detect QTLs in substitution lines was far to large. To define the position of genes on substitution chromosomes, recombinant inbred chromosome substitution lines can be developed (Law, 1966) and have been successful in the cloning of genes underlying traits in agriculture.

Whereas classical single strain and substitution line association studies had advantages in terms cost, coverage and reproducibility, their main weakness was a lack of

power for genome-wide association and low resolution. To obtain a higher resolution and an increase in power, we started to examine genetic reference panels (GRP). GRPs are defined as sets of individuals with fixed and known genomes that can be replicated indefinitely (i.e. inbred lines). Typically they consist of dozens to hundreds of inbred lines related by descent from a set of common ancestors (i.e., the founders). GRPs have been developed for many organisms, including yeast, plants, flies, and mammals (Crow, 2007; Buckler et al., 2009; Ayroles et al., 2009; Kover et al., 2009; Cubillos et al., 2011). While individual inbred lines are free of population structure, when considering a large number of inbred lines, there is evidence of population structure between the inbred lines (McClurg et al., 2007) complicating the analysis.

The Hybrid Mouse Diversity Panel (HMDP; Bennett et al., 2010) is an example of a large GRP which has been widely used. The HMDP increases the statistical power and resolution of the classical association studies by including a set of 70 recombinant inbred mouse strains in the mapping panel. In this design, approximately 100 strains are phenotyped (30 classical inbred strains and 70 recombinant inbred strains), and association is carried out after correcting for population structure using, for example, efficient mixed-model association (EMMA; Kang et al., 2008). By using the combined population included in the HMDP provides a high statistical power (from the recombinant inbred strains) and a high resolution (from the classical inbred strains). A limitation of the HMDP is the number of available inbred strains, resulting in an upper limit on the statistical power of the HMDP.

Rather than simply observing the differences in phenotypes between lines in a GRP, another option is to cross two inbred lines together. In a cross, two inbred strains are mated, and their offspring are either mated to each other (an intercross design) or to a progenitor strain (a backcross design). Second-generation offspring are then phenotyped and genotyped, and linkage analysis is carried out to identify a region that is associated

with the trait. However, each QTL region is large (i.e. low resolution), often containing tens of megabases and hundreds of genes. The process of identifying the causal variant and the gene involved is therefore difficult and costly. In a genetic cross, only a few hundred animals are required to identify loci that together explain 50% or more of the phenotypic variance for a particular trait. This finding is particularly striking compared to human studies, in which typically tens of thousands of individuals are required to identify loci that are involved in traits, and in which the loci identified typically explain only a small fraction of phenotypic variance (Flint and Eskin, 2012).

Another reference panel type approach is that of "in silico mapping" (Grupe et al., 2001). By "in silico" we mean a QTL mapping method which uses existing phenotypic and genotypic variation within common laboratory inbred strains for association studies. Over the years, breeding and inbreeding over the years has produced the commonly used modern laboratory strains of mice, and a wide variation of phenotypic traits have been observed (McClurg et al., 2006). The genotypic structure of these strains is also being explained through dense mapping of SNPs, and variance among these strains is emerging in the form of haplotype structure (Yalcin et al., 2004; Wiltshire et al., 2003). It was originally hypothesized that "in silico" mapping has the necessary experimental requirements to facilitate QTL mapping (Grupe et al., 2001; Pletcher et al., 2004), suggesting that phenotype-specific mouse crosses are not needed for the identification of QTL, and that large-scale genotyping efforts could be generated and combined in a phenotype-independent manner. While this may be true, many "in silico" mapping projects fell short of their goals due to the inability to properly assess the population structure prior to methods such as EMMA being adopted from techniques used in animal breeding where historically they have had to deal with related individuals.

The Collaborative Cross (CC) was proposed in 2002 as a large-scale multiparental recombinant inbred line panel as a project aimed at generating a common platform

for mammalian complex trait genetics that would overcome the limitations of existing resources (Threadgill, Hunter and Williams, 2002) and that can advance the field beyond complex trait analyses toward systems genetics (Threadgill, 2006). Unlike the HMDP, which consists of currently available strains, the Collaborative Cross has generated new inbred strains using a specific breeding scheme increasing power and resolution. The Collaborative Cross is also advantageous as there is less population structure than would be expected in a standard GRP. While techniques such as EMMA are available to correct for population structure, the presence of population structure still has a negative effect on statistical power. The final eight-way RIL design of the CC was community driven (Churchill et al., 2004) and included founders from five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/HlLtJ) and three wild-derived strains that were selected to represent three Musmusculus subspecies (CAST/EiJ, PWK/PhJ, andWSB/EiJ).

An alternative strategy to inbred lines is to use outbred mice. Bi-parental populations, or advanced intercross lines (AIL; proposed by Darvasi and Soller (1995)), have been used by selecting founder lines with large phenotypic differences for one or more traits, usually with unrelated parents selected to maximize marker polymorphisms (Doerge, 2002). They traditionally were difficult to analyze due to relatedness of individuals in the populations. With the introduction of methods to deal with the relatedness (see Sillanpää (2011) for a review of many options) the use of AILs and more complicated populations have recently become more widely used. These include heterogeneous stock mice (Demarest et al., 1999; Valdar et al., 2006) (for which animals are descended from eight classical inbred founder strains) and the Diversity Outbred (DO) mice (Svenson et al., 2012) (which comprises animals descended from the eight Collaborative Cross founder strains). Outbred mice can be viewed to be similar to F2 animals generated from a cross, but they have ancestry from eight founder strains instead of only two, and

the population is bred for more generations. The main advantage of HS/DO strategies is that they can be used to generate an almost limitless number of animals, enabling large studies to be carried out that can find weak genetic effects. In addition, owing to their breeding history, animals have undergone many more recombination events increasing mapping resolution.

1.3.2 Analysis of Outbreed populations

A number of experimental strategies have been proposed for association mapping of complex traits in model organisms. Many involve the use of highly recombinant populations derived from inbred lines. Examples of such populations are advanced intercross lines (AILs) , where a pair of inbred progenitors are intercrossed for three or more generations, and heterogeneous stocks (HS; Demarest et al., 1999), where a number, usually eight, of inbred strains are intercrossed for many generations. The Diversity outbreed (DO; Svenson et al., 2012) population has recently been developed in mice which resembles the HS in breeding structures. In theory, these strategies can achieve much higher-resolution mapping than that which is obtainable in standard inbred strain crosses. One such reason is they accumulate a greater density of recombinations, allowing for a finer mapping of the founders. Another issue is that the individuals in the population are related to some level, which often violates standard mapping techniques which may be applied to independent subjects.

Multiple founder recombinant populations have used similar breeding schemes to AILs (Valdar et al., 2006) but differ from AILs as they descend from more than two inbred strains, typically eight, adding additional complexity to the population. Because the markers used for genotyping will have fewer alleles than the number of haplotypes in the cross, individual markers typically do not unambiguously identify the underlying strain haplotype. In particular, unless all variants are genotyped, single-marker association analyses will fail to capture some QTL effects (Mott et al., 2000).

Polygenic based approaches

According to some recent views, population structure and relatedness between individuals both require their own correction terms (see Sillanpää (2011)), or may need additional correction after fitting a polygenic model (Amin, van Duijn and Aulchenko, 2007). Recently, linear mixed models have been shown to effectively correct for population structure in the association mapping of quantitative traits (Yu et al., 2006). Linear mixed models incorporate genetic relatedness between every pair of individuals directly as a random effect which addresses the correlation between individuals phenotypes due to their level of relatedness (e.g. siblings, first cousins, second cousins, etc.). This reflects the theory that the phenotypes of two genetically similar individuals are more likely to be correlated than those which are more dissimilar genetically. Applications of mixed models to association mapping in maize and potato panels demonstrate that mixed models obtain fewer false positives and higher power than previous methods including genomic control, structured association, and principal component analysis (Yu et al., 2006; Malosetti et al., 2007; Zhao et al., 2007).

Many highly recombinant model organism populations, such as the DO or HS, resemble those found in plant and animal breeding. Linear mixed models approach modeling the relatedness of individuals through variance components parameterized by the kinship matrix (Valdar et al., 2009). Specifically, the effects of a single locus are estimated simultaneously with one or more random intercept whose expected correlation structure is fixed given the kinship matrix based on the pedigree (or realized kinship matrix based on observed genotypes) and models the effects of overall genetic relatedness to account for effects from the rest of the genome (Kennedy, Quinton and Vanarendonk, 1992; Jannink, Bink and Jansen, 2001; Zhao et al., 2007).

This type of approach has been taken by two popular methods: Efficient Mixed-Model Association (EMMA; Kang et al., 2008) and QTLRel (Cheng et al., 2011).

EMMA was proposed as an efficient exact procedure that corrects for population structure and genetic relatedness in model organism association mapping during a period where it was not computationally efficient to use linear mixed effect models. EMMA takes advantage of the specific nature of the optimization problem in applying mixed models for association mapping, substantially increase computational speed and improved the reliability of results by achieving near global optimization (Kang et al., 2008). While this was a great improvement, the EMMA algorithm was still computationally infeasible for large data sets because the variance components parameters are estimated for each marker. A new implementation of the algorithm called Efficient Mixed-Model Association eXpedited (EMMAX; Kang et al., 2010) makes the simplifying assumption that because the effect of any given SNP on the trait is typically small, then the variance parameters only need to be estimated once for the entire dataset, rather than once for each marker. This change sacrificed the exact solution calculation from EMMA for a feasible computation time.

QTLRel (Cheng et al., 2011) is a more recent software which was developed to quickly perform genomewide scans, using a similar technique to EMMAX, with the advantage of having multiple random effects. While they specifically use the pedigree to infer the relationship matrix between individuals, this can be replaced by a realized kinship matrix based on the observed genotypes. One of the main advantages to QTLRel over EMMA is that it also has the ability to include other random effects such as cage effects, environment effects, or treatment effects.

While several approximate methods have been proposed address the issue of computation times of genomewide scans (e.g. EMMAX and QTLRel), efficient exact options exist. Zhou and Stephens (2012) propose an efficient exact method, which is refer to as genome-wide efficient mixed-model association (GEMMA) which makes approximations unnecessary in many contexts. The method is approximately n times faster than

the exact method EMMA and comparable to many approximate methods, making exact genome-wide association analysis computationally practical for large numbers of individuals. We note that in some settings the approximate methods provide results almost identical to those from the exact method (Kang et al., 2010; Zhang et al., 2010), it is not guaranteed in general.

Multiple locus based approaches

In a complex trait GWAS, the trait is affected by multiple functional loci and therefore a multiple locus association method would be preferred (Ayers and Cordell, 2010). To identify the important loci within the multiple locus model, variable selection or regularization of the predictors is required (e.g., Sillanpää and Bhattacharjee, 2005; Hoggart et al., 2008; O'Hara and Sillanpää, 2009; Wu et al., 2009; Ayers and Cordell, 2010; Cho et al., 2010).

The polygenic aspect of the model which accounts for both the distant (i.e., between populations) and close (i.e., within population) relatedness structures in the data can be addressed by a multiple locus model as the genetic relationships between the individuals can be captured by the markers themselves (e.g., Habier, Fernando and Dekkers, 2007). This allows for the possibility to use the models without additional polygenic terms. In Kärkkinen and Sillanpää (2012), they showed that multiple locus models that did not try to explicitly model polygenic effects worked well. Their observation of the redundancy in including additional polygenic components is in agreement with, for example, Calus and Veerkamp (2007) and Pikkuhookana and Sillanpää (2009). Furthermore, Calus and Veerkamp (2007) claim that including polygenic effects at higher SNP densities will not improve the accuracy of total breeding values. Specifically, they found that when the average LD, measured as r^2, between adjacent markers is at least 0.10, depending on the heritability of the trait, there appears to be little reason to include a polygenic effect in the model.

Utz, Melchinger and Schön (2000) implement a multiple locus resampling based procedure for detecting functional loci in GWAS, and showed in their simulations that the resampling was able to correct some biases and sampling errors in the model estimation. Schön et al. (2004) used composite interval mapping by the regression approach (Haley and Knott, 1992) in combination with the use resampling of an multiple locus additive genetic model, as done in Utz, Melchinger and Schön (2000) with loci selected by stepwise regression for the analysis of test cross progenies. They found that for even moderate sample sizes that their procedure was able to obtain estimates with very low bias. They concluded that for traits regulated by a few QTL with large effects, for which phenotypic selection is expensive or hampered due to rare occurrence, that resampling multiple locus approach of MAS (Utz, Melchinger and Schön, 2000) can be very useful.

Another resampling based multiple locus method called frequentist model averaging (FMA) was proposed in Hjort and Claeskens (2003). FMA examines each combination of predictors multiple locus models and averages over the models with weights to obtain parameter estimates. FMA can be implemented without much difficulty or protracted computations. One requirement of FMA is the specification of model weights. Several method to define the weights have been proposed which include AIC weights (Buckland, Burnham and Augustin, 1997), weights based on minimizing a Mallows criterion (Hansen, 2007), and weights based on the Focused Information Criterion (Claeskens and Consentino, 2008). Williams and Christian (2006) showed that FMA estimates for genetic effects in twins studies were more accurate than the standard estimates based on the criteria used for the model averaging weights. Schomaker, Wan and Heumann (2010) address the issue of missing data in the FMA framework. They proposed how one can incorporate imputation first and then preform FMA rather than attempt to incorporate complex weighting adjustments to criteria such as AIC which allow for

missing data (e.g., the EM-based AIC developed in Claeskens and Consentino (2008)). They also propose a frequentist model selection (FMS) estimator which is a special case of FMA which focuses on the selected model rather than the estimated effects.

The QTLMAS XII meeting provided a common data set for which attendees could propose methods to analyze the data. The summaries of submitted methods support recent views for of a preference for multiple locus models (Crooks et al., 2009). The results from LDHap (Ledur, Navarro and Pérez-Enciso, 2009) were best overall in this dataset, with LABayes (Bink and van Eeuwijk, 2009) and LDBayes (Cleveland and Deeb, 2009) having the second highest power for QTL detection. As LDHap and LABayes both used information from several markers for detecting QTLs, it suggests that multiple marker methods may have higher power to find QTLs.

Other approaches

Although polygenic effects and multiple locus modeling are popular methods, other methods have been proposed. Other widely used methods for related individuals in human association mapping include genomic control (Devlin and Roeder, 1999), structured association (Pritchard, Stephens and Donnelly, 2000), and principal component analysis (Patterson, Price and Reich, 2006; Price et al., 2006). However, these methods have shown to be inadequate within the realm of model organisms. Genomic control has reduced power when the effect of population structure is large, as would be expected in model organisms (Yu et al., 2006). Principal component based analyses, which assume only a small number of ancestral populations and admixture, are only able to partially capture the multiple levels of population structure and genetic relatedness in model organisms (Aranzana et al., 2005; Yu et al., 2006; Zhao et al., 2007).

1.4 Coding of alleles, and modeling effects

1.4.1 SNP effects

When defining a regression model with genetic predictors such as SNPs that are not quantitative in nature, the way they are defined can have a large effect on the model. As previously mentioned, it is common practice in statistical genetics to code SNPs by the count of the minor allele Q as {0,1,2} for the unphased genotypes {qq, qQ, QQ}.

Additive SNP effects

The most common genetic model is a simple additive model. Under the additive model no modifications need to be made from the {0,1,2} SNP coding as the count of the minor allele is a natural predictor for an additive effect for the presence of the allele.

Dominant SNP effects

When one is interested in accounting for dominance in the regression model, the standard {0,1,2} SNP coding is unable to capture the dominant aspect of the effect. In order to model dominance, we need to include more than one predictor in the model for each SNP, i.e. two predictors for single locus regressions and $2p$ predictors for the multiple locus model with p loci.

There are two standard methods for encoding dominant predictors for regression models. The first is to keep the standard {0,1,2} additive predictor to account for the additive nature in the dominance effect, and add a second predictor to account the deviation from additivity. This second predictor will be an indicator variable for the unphased genotype qQ. That is, if we denote g_i as the number of the minor allele in the genotype, that is the standard {0,1,2} coding of a genotype, then we would have an additive predictor, $a_i = g_i$, and a dominant predictor, $d_i = I(g_i = 1)$.

The second method to encode dominance into the regression model is to ignore the standard additive predictor and define two indicator variables. The first indicator is for

the effect of having one copy of the minor allele, $d_i^1 = I(g_i = 1)$, and a second indicator for the effect of having two copies of the minor allele, $d_i^2 = I(g_i = 2)$. Ths is equivalent to an ANOVA model. Both modeling variations have proven useful in detecting dominant effects. With either coding, there exists a high potential for collinearity in the model, especially for SNPs with low minor allele frequencies (MAF).

1.4.2 Haplotype effects

Rather than the traditional observed marker data (e.g. SNPs), we are interested in modeling the subjects haplotypes in the intervals between observed markers. In the context of multiple founder crosses, we can use the detailed founder haplotype information to identify the state of each subject in the interval. In brief, haplotype descent along each subject's genome can be inferred by the haplotype reconstruction method HAPPY (Mott et al., 2000), which applies a hidden Markov model simultaneously to the genotypes of the founder strains and the n subjects. For each subject, at each interval, i.e. between adjacent pairs of observed markers, HAPPY produces a vector $\boldsymbol{g}_i(m)$ containing the descent information from the founders at marker m based on either an additive or full effect model which are described in detail below.

Before describing the exact form of $\boldsymbol{g}_i(m)$, let us consider a cross with J founders. For each locus, the subject within the cross will have two haplotypes present, one on each copy of the chromosome which the locus presides. For each individual, HAPPY will provide us with a $J \times J$ matrix $\mathbf{P}$, where p_{ij} is the probability that the first haplotype is from founder i and the second haplotype is from founder j. We summarize $\mathbf{P}$ as $\boldsymbol{g}(m)$ based on the selected model described below.

Additive Model

The additive haplotype model describes the locus based on the expected number of each founders haplotype present at each given locus. For a J founder cross, the additive version of $\boldsymbol{g}(m)$ is a J-vector and which sums to 2. The exact definition of the additive

locus predictor for subject i at locus m is given by

$$\boldsymbol{g}_i^a(m) = \mathbf{1}^{\mathrm{T}}(\mathbf{P} + \mathbf{P}^{\mathrm{T}}), \tag{1.2}$$

where $g_{i,j}^a = E(\text{number of haplotype } j)$ and $\mathbf{1}^{\mathrm{T}}\mathbf{g}^a = 2$

Full Model

The full diplotype model describes the locus based on the probability of each unique founder haplotype pair (or diplotype). For a J founder cross, the full model version of $\boldsymbol{g}(m)$ is a $(J(J-1)/2)$ length probability vector. The exact definition of the full model locus predictor for subject i at locus m is given by

$$\boldsymbol{g}_i^f(m) = \text{vech}\left(\mathbf{P} + \mathbf{P}^{\mathrm{T}} - \text{diag}(\text{vecdiag}(\mathbf{P}))\right), \tag{1.3}$$

where vech() returns the upper triangle matrix, including the diagonal, as a vector, vecdiag() returns the diagonal as a vector, and $\mathbf{1}^{\mathrm{T}}\mathbf{g}^f$ =1.

1.5 Overview of method comparison with ROC curves

Performance of a method may be evaluated formally using receiver operator characteristic (ROC) curves. ROC curve methodology can vary between studies (Krzanowski and Hand, 2009), so we describe ours in full. A given simulation study comprises a set of simulation trials $\mathcal{S} = \{1, \ldots, S\}$. In each trial s, a given method is presented with m SNPs of which m_q will be causal. That method calculates a single score for each SNP (an RMIP or logP). For a given threshold t, define $\text{power}_s(t)$ as the proportion of m_q causal SNPs scoring $\geq t$ (ie, the power to detect), and the false positive rate $\text{FPR}_s(t)$ as the proportion of $m - m_q$ non-causal SNPs scoring $\geq t$. We define the area under curve in trial s for FPRs between a and b as $\text{AUC}_s(a, b) = \int_a^b \text{power}_s\left(\text{FPR}_s^{-1}(x)\right)\, dx$,

where $\text{FPR}_s^{-1}(x)$ returns the threshold t at which the FPR is x, and the integration is approximated using the trapezoid rule. For a given method and set of simulations $\mathcal{S}$ we define the estimated AUC between FPR a and b as $\overline{\text{AUC}}(a,b) = \sum_{s\in\mathcal{S}} S^{-1}\text{AUC}_s(a,b)$, and by the central limit theorem this estimate is approximately normally distributed with variance $(S-1)^{-1}\sum_{s\in\mathcal{S}}\left(\overline{\text{AUC}}(a,b) - \text{AUC}_s(a,b)\right)^2$. We define the "initial ROC" as the ROC curve in the range FPR $\in [0, 0.05]$, and the "initial AUC" as $\overline{\text{AUC}}(0, 0.05)$; the "full ROC" is where FPR $\in [0,1]$ and the "full AUC" is $\overline{\text{AUC}}(0,1)$. When plotting ROC curves for each method we use threshold averaging (Fawcett, 2006), varying t over its range ($[0,1]$ for RMIPs; $[0,\infty)$ for logP) and at each t plotting x and y coordinates $S^{-1}\sum_{s\in\mathcal{S}}\text{FPR}_s(t)$ and $S^{-1}\sum_{s\in\mathcal{S}}\text{power}_s(t)$ respectively.

1.6 Statistical questions of interest

The main statistical genetics problems that will be addressed in this proposal are:

- How does one detect genetic loci that effect a complex phenotype's expected value? This question will be addressed at multiple levels. We will first restrict ourselves to loci where SNPs are present and consider how to detect the SNPs that have a true effect under the following model assumptions:
  - The effects on the phenotype are additive. (See Chapter 2)
  - Allowing for general SNP effects, e.g,. dominance. (See Section 3)
- How does one detect genetic loci that effect a complex phenotype's mean value when the loci of interest are not SNPs?
  - The case we will consider is when the predictors are a single locus, but a group of variables must be used to model the locus, e.g. haplotypes. (See Chapter 4)

- How can Resample Model Averaging be used to gain information about variable relationships relevant to a particular phenotype? (See Chapter 6)

Chapter 2

Resample Model Averaging with the LASSO: LLARRMA

This chapter describes **L**ASSO **L**ocal **A**utomatic **R**egularization **R**esample **M**odel **A**veraging (LLARRMA; Valdar et al., 2012), a tool used to identify the predictors in a generalized linear model that are actually associated with a complex response out of a large set of potential predictors. The method was developed to assist in loci selection in GWAS type analyses.

This chapter is laid out as follows. Section 2.1 discusses the motivation behind developing LLARRMA. Section 2.2 describes the LLARRMA procedure in a general GLM setting and in the setting of a GWAS "hit region" analysis. Section 2.3 describes the simulation models that LLARRMA has been tested under. Section 2.4 compares the performance of LLARRMA with the traditional single locus regression methods and stability selection. Section 2.5 discusses the performance of LLARRMA and the advantages of LLARRMA.

2.1 Motivation

Single locus regression has become a staple tool of human genome wide association studies (GWAS; WTCCC, 2007). Despite the fact that it simplistically reduces the often complex genetic architecture of a phenotype down to effects at an individual

SNP (or other localized variant), it has proved powerful in identifying major genetic determinants and predictors of disease susceptibility (Cantor, Lange and Sinsheimer, 2010). Many would acknowledge that simultaneous modeling of all loci potentially yields fairer estimates of genetic effects, more stable phenotypic predictions, and better characterization of between-locus confounding (Lee et al., 2008; Hoggart et al., 2008). However, such multiple locus approaches are at present seldom used. This could be because they are considered impractical, too abstruse, or, with some theoretical support (Fan and Lv, 2008), unnecessary for an initial genome scan. Certainly, much of the genomewide confounding that explicit multiple locus modeling would hope to resolve is efficiently, if bluntly, dealt with by the addition of regression covariates correcting for higher order geometric relationships in the data (Price et al., 2010), or demographically or probabilistically inferred strata (Pritchard, Stephens and Donnelly, 2000).

Nonetheless, once initial genome scans have been performed and "hit regions" of association identified, the short-comings of a single locus approach become painfully apparent. Local patterns of linkage disequilibrium (LD) in such hit regions can make both the number of underlying causal signals and the identity of the loci that most directly give rise to them (eg, Strange et al., 2010) ambiguous. Statistical analysis after this point is often of an ad hoc nature. It typically involves fitting further regressions that condition on "top" loci that appear most strongly associated in order to rule out neighbors or rule in suspicions of an independent second signal. This is quickly followed by more interpretive analysis based on annotation as a prelude to, for example, investigation at the bench. In ad-hoc conditioning, rarely are there formal consideration of the fact that the association of the top locus is often insignificantly different from that of its correlated neighbors, and that whereas its association with the phenotype is probably stable to sampling error, its superiority in association over its neighbors

is probably not. This inherent instability of the relative strengths of association between confounding loci makes such strategies extremely high risk: a slightly different sampling of individuals could demote the conditioning locus, resulting in an alternative conditioning locus and potentially lead to drastically altered conclusions. This approach becomes more precarious still when some of the loci are themselves known with varying certainty, their genotypes having been partially or wholly imputed (Zheng et al., 2011), such that weakness of association is now also a function of imputation uncertainty unrelated to the phenotype (eg, Servin and Stephens, 2007).

There is thus great value in developing principled approaches to discriminate true from false associations in hit regions. Joint modeling of all loci through multiple regression seems attractive because it accounts for the LD of the data (Balding, 2006). However, while many approaches to the multiple locus modeling have been attempted, they all have a common approach to the problem. Whether it be penalized regression (eg, Wu et al., 2009; Zhou et al., 2011), Bayesian regression (eg, Balding, 2006), or resampling methods (eg, Valdar et al., 2009; Alexander and Lange, 2011; He and Lin, 2011), they all focus on a genome-wide perspective for SNP selection. While the methods implemented provide innovative approaches to the problem, attempting to not only select hit regions from a genome-wide perspective but also handling localized LD structure is often too hard of a task.

We propose a different approach. Rather than attempting to handle two vastly different problems with the same method, we assume that the selection of hit regions of association can be handled sufficiently well by, for example, standard single locus regression, and focus on the more complex problem of handling localized LD structure. Prescreening by, for example, single locus regression is motivated by SIS (Fan and Lv, 2008), but rather than just considering the most significant SNPs, we propose to analyze the entire region of LD identified by the top SNPs. While methods such as

GWASelect (He and Lin, 2011) directly incorporate an iterative SIS (Fan and Lv, 2008) approach to ensure that the selection of only the top SNPs has not missed important SNPs. By simply selecting the entire hit region rather than just the top SNPs, we avoid additional computational expenses from the iterative approach.

2.2 Methods

We start by considering a generalized linear model (GLM) to estimate the effects of m variables on an outcome with n individuals, and then describe statistical approaches to identify a subset m_q of variables that are be truly influential. We assume that the m variables may be highly correlated, and that $m_q < m < n$. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be an n-vector of responses, let $\mathbf{X}$ be an $n \times m$ matrix of generic predictors, and let $\mathcal{D} = \{\mathbf{y}, \mathbf{X}\}$ and $\mathcal{N} = \{1, \ldots, n\}$. The GLM models a link function of the responses by a linear function of the m SNP predictors

$$f(\eta_i) = \mu + \sum_{j=1}^{m} \beta_j x_{ij} \tag{2.1}$$

where $\eta_i = E(y_i)$, $f(\cdot)$ is a link function, and β is a vector of regression coefficients.

We assume that only a subset of the m predictors have a genuine effect on the response, and define a corresponding indicator vector of inclusions $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ such that $\gamma_j = I(\beta_j \neq 0)$. A common way to infer $\boldsymbol{\gamma}$, and to thereby identify the set of truly influential predictors, is to use a model selection procedure that maximizes some criterion of model fit. This returns a binary vector $\hat{\boldsymbol{\gamma}}$, an estimate declaring which predictors belong in the model. Although attractive, $\hat{\boldsymbol{\gamma}}$ has limited interpretability because it provides no information about how sensitive the selection could have been to finite sampling. That is, would $\hat{\boldsymbol{\gamma}}$ be expected to vary dramatically when applied to alternative samples from the same population. Moreover, although many selection

procedures are consistent, this provides little reassurance when the sample is finite, and suggests that the returned statistic $\hat{\boldsymbol{\gamma}}$ could have high variance.

2.2.1 General Framework

While LLARRMA was developed for the application of genetic association studies, the method can be applied more generally. To emphasize the generality of the method, we will first provide all details of LLARRMA in its most general form. We note that while the method has only been implemented for GLMs with Gaussian or binomial responses, with their corresponding canonical link functions (i.e., standard linear regression or logistic regression), it can be implemented for other GLMs with proper modifications to the formulas presented below.

Resample model averaging

We seek to estimate $\boldsymbol{\gamma}$ in way that incorporates uncertainty in model choice. An example of such uncertainties is the potential variability of the selected model due to finite sampling. To do this, we use resample model averaging (RMA; Valdar et al., 2009), applying a model selection procedure to repeated resamples of the data, and base subsequent inference on the aggregate of those results. While the general RMA procedure proposed by Valdar et al. (2009) provides the choice of either subsampling or bootstrap samples, we have selected to use subsampling (detailed justification provided later). Rather than obtaining a binary estimate of each γ_j, we instead seek to estimate its expectation $\mathrm{E}(\gamma_j)$ over resamples, hoping to approximate its expectation over samples from the population. We start by drawing subsamples $k = 1, \ldots, K$ with subsampling proportion $\phi = \frac{2}{3}$, such that each subsample comprises data $\mathcal{D}^{(k)} = \{\mathbf{y}^{(k)}, \mathbf{X}^{(k)}\}$ on $|\mathcal{N}^{(k)}| = \phi n$ individuals such that $\mathcal{N}^{(k)} \subset \mathcal{N}$. Each subsample is produced by drawing ϕn individuals at random without replacement. For each subsample k, we perform a fixed model selection procedure to estimate $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)}) = \hat{\boldsymbol{\gamma}}^{(k)}$, the m-length indicator vector of model inclusions based on the kth subsample. Applying this to all subsamples

gives the $K \times m$ matrix $\mathbf{\Gamma}$, where $\mathbf{\Gamma}^{\mathrm{T}} = [\hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\gamma}}^{(2)}, \cdots, \hat{\boldsymbol{\gamma}}^{(K)}]$. The expected proportion of times that the jth predictor is included in the model is given by its RMA estimate

$$\widehat{\mathrm{RMIP}}_j = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)})_j = \frac{1}{K}\sum_{k=1}^{K} \hat{\gamma}_j^{(k)} = \frac{1}{K}\sum_{k=1}^{K} \Gamma_{jk}\,, \tag{2.2}$$

which we refer to as its resample model inclusion probability (RMIP).

RMA - use of subsampling over bootstrap sampling

Sample aggregation techniques such as bootstrap aggregation (referred to as bagging (Breiman, 1996)) or subsample aggregation (referred to as subagging (Bühlmann and Yu, 2002)) have been found to be useful when estimating an indicator parameters such as $\boldsymbol{\gamma}$. We have chosen to use subsample aggregation, or subagging, in LLARRMA. The choice to use subagging was influenced by similar choices of Valdar et al. (2009) and stability selection (Meinshausen and Bühlmann, 2010), along with the theoretical justification and primary application related reasons discussed below.

The results of Politis, Romano and Wolf (1999, p. 47-51) comparing the bootstrap and subsampling procedures provide some theoretical justification for our use of subsampling over bootstrapping in this setting. Specifically, Politis, Romano and Wolf (1999) discuss that bootstrap methods most often require the assumption that the distribution of the estimated statistic is at least locally smooth, an assumption that is not needed for subsampling. As the true distribution of γ, an indicator function, is not smooth, the results were important in our choice of subsampling.

We also have chosen to use subsampling rater than bootstrapping for reasons specific to our primary application. Specifically, in the setting of a genetic association study, resampling with replacement does not properly fit the genetic model. By this, we mean that the probability that we observe multiple individual with the same genetic composition is very rare, yet in a bootstrap sample it is a very common event.

Model selection within a subsample using the LASSO

To select variables within the kth subsample we use LASSO penalized regression (Tibshirani, 1996). This estimates $\boldsymbol{\beta}$ for subsample k as

$$\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)}) + \lambda \sum_{j=1}^{m} |\beta_j| \right\}, \tag{2.3}$$

where $\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)})$ is the log-likelihood of $\boldsymbol{\beta}$ for data $\mathcal{D}^{(k)}$, and λ is a penalty parameter. The LASSO estimate $\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)})$ easily translates into an estimate of the inclusions $\hat{\boldsymbol{\gamma}}(\lambda; \mathcal{D}^{(k)}) = I(\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) \neq 0)$. However, to arrive at the single estimate of $\boldsymbol{\gamma}$ required for the RMA procedure, we must devise a suitable criterion for choosing the penalty parameter λ. We propose two alternatives, both of which identify a value $\lambda^{(k)}$ specific to subsample k (ie, local): complement deviance selection and permutation selection.

Predictive-based choice of $\lambda^{(k)}$: complement deviance selection

The complement deviance criterion seeks a model that would perform well in out-of-sample prediction. After estimating $\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)})$ over a grid of λ when calculating the LASSO regression path, this criterion finds the value of λ that minimizes the deviance of the complement of subsample k, ie,

$$\hat{\lambda}^{(k)}_{\text{CompDev}} = \underset{\lambda}{\operatorname{argmin}} \text{ deviance}(\mathcal{D}^{(\setminus k)}, \lambda),$$

where $\mathcal{D}^{(\setminus k)}$ is the data of the $(1-\phi)n$ individuals not selected for subsample k, and deviance$(\mathcal{D}^{(\setminus k)}, \lambda)$ is the deviance of data $\mathcal{D}^{(\setminus k)}$ from the model fit on data $\mathcal{D}^{(k)}$ with penalty parameter λ.

Discovery-based choice of $\lambda^{(k)}$: permutation selection

The permutation selection criterion is a modified version of that proposed by Ayers and Cordell (2010) and seeks a conservative model that would tend to include no

varaibles under permutation of the response. Given a subsample k, we estimate for a given permutation of the response $\boldsymbol{\pi}(\mathbf{y})$ the smallest penalty required to zero out all predictors, $\lambda_{\text{null}}(\boldsymbol{\pi}, k)$, a formula that will depend on the GLM model used. Calculating this for each of S permutations $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S$, we estimate the permutation selection λ for subsample k as

$$\hat{\lambda}^{(k)}_{\text{Perm}} = \text{median}(\{\lambda_{\text{null}}(\boldsymbol{\pi}_1, k), \lambda_{\text{null}}(\boldsymbol{\pi}_2, k), \ldots, \lambda_{\text{null}}(\boldsymbol{\pi}_S, k)\}) \,. \tag{2.4}$$

Ayers and Cordell (2010) apply a similar criterion when analyzing complete data sets, with the difference that they estimate $\hat{\lambda}_{\text{null}}$ as the maximum of $\{\lambda_{\text{null}}(\boldsymbol{\pi}_1), \ldots, \lambda_{\text{null}}(\boldsymbol{\pi}_S)\}$ for $S = 25$. However, we prefer not to do this for two reasons. The first is that the maximum is relatively unstable statistic for $S = 25$, and secondly the max is undesirable, especially for larger S, since it potentially allows $\hat{\lambda}_{\text{null}} = \lambda_{\text{null}}(\boldsymbol{\pi}_s)$ where $\boldsymbol{\pi}_s(\mathbf{y}) = \mathbf{y}$, an improper null permutation as this is a permutations assumed to be in the alternative state. In contrast, when using the median (Eq 2.4) the accuracy of $\hat{\lambda}_{\text{Perm}}$ increases with S, although we find that in simulations $S = 20$ is adequate.

Incorporating uncertainty due to missing data: single and multiple imputation

Data often includes combinations of variables and subjects for which the data is unknown or uncertain. To avoid a potentially wasteful complete cases analysis, it is common to impute the missing data and analyze the partly-imputed data as if it were fully observed. Dividing the data matrix $\mathbf{X}$ into missing and observed elements $\mathbf{X} = \{\mathcal{X}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$, imputation methods based on bayesian posterior distributions model the joint distribution $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ includes additional information used in the imputation (e.g., priors). Most studies, however, do not use this joint distribution directly. Rather, they replace $\mathcal{X}_{\text{mis}}$ with a point estimate $\widehat{\mathcal{X}}_{\text{mis}}$, each element of which is constructed from it's marginal distributions. Specifically, $\mathcal{X}_{\text{mis}}$ is replaced by a "hard"

imputation, $\widehat{\mathcal{X}}_{\text{mis}}^{\text{hard}}$, with elements imputed as their maximum a posteriori value

$$\hat{x}_{ij} = \underset{g \in \mathcal{G}}{\text{argmax}}\ p(x_{ij} = g | \mathcal{X}^{\text{obs}}, \boldsymbol{\omega}).$$

where $\mathcal{G}$ is the set of all posible values x_{ij} may take.

The simplest approach to modeling missing data within LLARRMA is first to estimate $\mathcal{X}_{\text{mis}}$ as $\widehat{\mathcal{X}}_{\text{mis}}^{\text{hard}}$ and then subsample $\widehat{\mathbf{X}} = \{\widehat{\mathcal{X}}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$ as if it were complete. This plug-in approach underestimates variability because it fails to incorporate uncertainty about the imputation. However, ignoring imputation uncertainty could be more problematic in multiple locus settings, if, for example, the posterior distribution of the data $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ differs substantially from joint distribution implied by the product of marginal posteriors $\prod_{ij \in \mathcal{X}_{\text{mis}}} p(x_{ij}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ (eg, Servin and Stephens, 2007). A natural way to incorporate imputation uncertainty into our resampling framework is through multiple imputation (Little and Rubin, 2002). At each iteration k, we sample a new $\mathcal{X}_{\text{mis}}^{\star}$ from its posterior $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$, subsample the resulting $\mathbf{X}^{\star} = \{\mathcal{X}_{\text{mis}}^{\star}, \mathcal{X}_{\text{obs}}\}$ to give $\{\mathbf{X}^{\star(k)}, \mathbf{y}^{(k)}\} = \mathcal{D}^{\star(k)}$, and then calculate RMIPs using $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{\star(k)})$ in place of $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)})$ in Eq 2.2. The resulting RMIPs incorporate additional variability because each subsample now includes a potentially different imputation of the missing data. We implement both hard and multiple imputation options for LLARRMA as there may not always be a multiple imputation method available for a given data type.

2.2.2 Implementation for genetic association studies

The data model

For our primary application of LLARRMA, we start by considering a standard logistic regression to estimate the effects of m SNPs in a hit region on a case/control outcome in n individuals, and then describe statistical approaches to identify a subset m_q of SNPs that might be truly influential. We assume that the data is derived from a hit region

that has been previously identified by an initial genomewide screen using, for example, single locus regression, and that the m SNPs may be highly correlated with each other due to blocks of LD. We assume $\mathbf{y} = (y_1, \ldots, y_n)$ to be an n-vector of the dichotomous response with each of the n_1 cases coded by 1 and the n_0 controls coded by 0, and that $\mathbf{X}$ is an $n \times m$ matrix of SNP genotypes, where SNPs are coded to reflect additive-only effects as $\{0, 1, 2\}$ for unphased genotypes {qq,qQ,QQ}. Logistic regression models the case-control status of individual i as if sampled from $Y_i \sim \text{Bin}(p_i, 1)$, where i's propensity $p_i = P(Y_i = 1)$ is determined by a linear function of the m SNP predictors

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mu + \sum_{j=1}^{m} \beta_j x_{ij}\,, \tag{2.5}$$

where x_{ij} is value of the jth SNP for the ith individual and the ijth element of the column-centered design matrix $\mathbf{X}$, μ is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ are the effects of the m predictors.

Resample model averaging

The RMA procedure described above does not need any modification or further clarification for the genetic association hit region application.

Selection within a subsample using the LASSO

To select SNPs within the kth subsample we use LASSO penalized logistic regression. This estimates $\boldsymbol{\beta}$ for subsample k as

$$\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ -\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)}) + \lambda \sum_{j=1}^{m} |\beta_j| \right\}, \tag{2.6}$$

where $\ell(\boldsymbol{\beta}; \mathcal{D}^{(k)})$ is the log-likelihood of $\boldsymbol{\beta}$ for subsampled data $\mathcal{D}^{(k)}$, and λ is a penalty parameter. The LASSO estimate $\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)})$ still translates into an estimate of the inclusions $\hat{\boldsymbol{\gamma}}(\lambda; \mathcal{D}^{(k)}) = I(\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)}) \neq 0)$ as before.

Predictive-based choice of $\lambda^{(k)}$: complement deviance selection

The complement deviance criterion described in the general setting may be explicitly defined for logistic regression. After estimating $\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)})$ over a grid of λ in order to calculate the LASSO path, this criterion finds the value of λ that minimizes the deviance of the complement of subsample k, ie,

$$\hat{\lambda}^{(k)}_{\text{CompDev}} = \underset{\lambda}{\text{argmin}} \left\{ -2 \sum_{i \in \mathcal{N}^{(\backslash k)}} [y_i \log(\hat{p}_{i,\lambda}) + (1 - y_i) \log(1 - \hat{p}_{i,\lambda})] \right\},$$

where $\mathcal{N}^{(\backslash k)} = \mathcal{N} \backslash \mathcal{N}^{(k)}$ is the set of $(1 - \phi)n$ individuals not selected for subsample k, and $\hat{p}_{i,\lambda}$ is the predicted probability of $P(Y_i = 1)$ based upon $\hat{\boldsymbol{\beta}}(\lambda; \mathcal{D}^{(k)})$ applied to the design matrix of the complement subsample $\mathbf{X}^{(\backslash k)}$.

Discovery-based choice of $\lambda^{(k)}$: permutation selection

The permutation selection criterion described in the general setting can be explicitly defined for logistic regression. Given a subsample k, we calculate for a given permutation of the response $\boldsymbol{\pi}(\mathbf{y})$ the smallest penalty required to zero out all predictors is given by

$$\lambda_{\text{null}}(\boldsymbol{\pi}, k) = \frac{1}{|\mathcal{N}^{(k)}|} \max_j \left| \langle \mathbf{x}^{(k)}_j, \boldsymbol{\pi}(\mathbf{y}^{(k)}) \rangle \right|$$

where $\mathbf{x}^{(k)}_j$ is the jth column of the subsampled and mean-centered design matrix $\mathbf{X}^{(k)}$, and $\langle \cdot, \cdot \rangle$ denotes the inner product of its two arguments. Calculating this for each of S permutations $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S$, we estimate the permutation selection λ for subsample k as given by Equation 2.4.

Incorporating uncertainty due to missing genotypes: hard, dosage and multiple imputation

SNP data within a hit region will often include combinations of markers and individuals for which the genotype is unknown or uncertain. To avoid a potentially wasteful complete cases analysis, it is common to impute the missing genotypes using a program

such as MACH (Li et al., 2010), IMPUTE (Howie, Donnelly and Marchini, 2009) or fastPHASE (Scheet and Stephens, 2006), and analyze the partly-imputed data as if it were fully observed. Imputation methods are typically based on reconstruction and phasing of inferred haplotypes. Dividing the SNP matrix $\mathbf{X}$ into missing and observed elements $\mathbf{X} = \{\mathcal{X}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$, methods such as fastPHASE (Scheet and Stephens, 2006) model the joint distribution $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ includes additional information used in the imputation (eg, priors). Most GWAS studies, however, do not use this joint distribution directly. Rather, they replace $\mathcal{X}_{\text{mis}}$ with a point estimate $\widehat{\mathcal{X}}_{\text{mis}}$, each element of which is constructed from its marginal distributions. Specifically, $\mathcal{X}_{\text{mis}}$ is replaced by either the "dosage", $\widehat{\mathcal{X}}^{\text{dose}}_{\text{mis}}$, with elements defined as the expectation of the allele count $\hat{x}_{ij} = \text{E}(x_{ij}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$; or a "hard" imputation, $\widehat{\mathcal{X}}^{\text{hard}}_{\text{mis}}$, with elements imputed as their maximum a posteriori genotype

$$\hat{x}_{ij} = \underset{g\in\{0,1,2\}}{\text{argmax}}\ p(x_{ij} = g|\mathcal{X}^{\text{obs}}, \boldsymbol{\omega}).$$

The simplest approach to modeling missing genotypes within LLARRMA is first to estimate $\mathcal{X}_{\text{mis}}$ as either $\widehat{\mathcal{X}}^{\text{dose}}_{\text{mis}}$ or $\widehat{\mathcal{X}}^{\text{hard}}_{\text{mis}}$ and then subsample $\widehat{\mathbf{X}} = \{\widehat{\mathcal{X}}_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$ as if it were complete. This plug-in approach underestimates variability because it fails to incorporate uncertainty about the imputation. Zheng et al. (2011) show that doing this when modeling effects at single loci reduces power a negligible amount when the imputation accuracy is reasonably high. Nonetheless, ignoring imputation uncertainty could be more problematic in multiple locus settings, if, for example, the posterior distribution of haplotypes $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ differs substantially from joint distribution implied by the product of marginal posteriors $\prod_{ij\in\mathcal{X}_{\text{mis}}} p(x_{ij}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$ (eg, Servin and Stephens, 2007). A natural way to incorporate imputation uncertainty into our re-sampling framework is through multiple imputation (Little and Rubin, 2002). At each iteration k, we sample a new $\mathcal{X}^{\star}_{\text{mis}}$ from its posterior $p(\mathcal{X}_{\text{mis}}|\mathcal{X}_{\text{obs}}, \boldsymbol{\omega})$, subsample the

resulting $\mathbf{X}^\star = \{\mathcal{X}^\star_{\text{mis}}, \mathcal{X}_{\text{obs}}\}$ to give $\{\mathbf{X}^{\star(k)}, \mathbf{y}^{(k)}\} = \mathcal{D}^{\star(k)}$, and then calculate RMIPs using $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{\star(k)})$ in place of $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)})$ in Eq 2.2. The resulting RMIPs incorporate additional variability because each subsample now includes a potentially different imputation of missing genotypes. We implement hard, dosage and multiple imputation using posterior draws from fastPHASE (making use of the -s option).

2.3 Simulation Framework

2.3.1 Simulation study 1: 5 loci in Cancer data

We obtained genotype data from phase 1 of a case-control GWAS for colorectal cancer from collaborators at the Wellcome Trust Centre for Human Genetics, University of Oxford. Two forms of the data are used here. The "cancer data" comprises complete genotype information on 1493 subjects for 183 SNPs covering a hit region previously identified on 18q21. The cancer data is a subset of the "full cancer data", which comprises incomplete genotype information on 1859 subjects for the hit region.

Generating missing genotypes

To assess the sensitivity of the compared methods to alternative strategies for modeling missing genotypes, we generate incomplete versions of the cancer data by deleting genotypes according to a random missingness algorithm. The missingness algorithm is based on empirical modeling of the pattern of missing data in the full cancer data. The full cancer data genotypes contained 854 missing genotypes ($\sim$0.25%). We observed that the proportion of missing genotypes varied considerably from SNP to SNP, but that missingness across individuals was consistent with a random allocation. To generate each incomplete data set, we therefore do the following. First, for each SNP j, we assign a missingness proportion $\psi_{\text{mis},j}$ generated as a random draw $\psi_{\text{mis},j} \sim f_{\text{mis}}$, where f_{mis} is an empirical density based on the histogram of missingness proportions of SNPs in the full cancer data. Second, we select a subset of size $n_{\text{mis}} < n$ individuals eligible to receive

missing genotypes. Third, at each SNP j we delete $d_j = n_{\text{mis}} \times \min(c\psi_{\text{mis},j}, 1)$ marker genotypes at random from the n_{mis} individuals which may be given missing values, where c is chosen such that the overall proportion of missing data, $p_{\text{mis}} = (mn)^{-1} \sum_j d_j$, is a fixed value. To generate a more conservative level of missingness while ensuring at least 10% of individuals had complete data, we set $p_{\text{mis}} = 0.118 \pm 0.0125$ and $n_{\text{mis}} = 0.9n$

Simulating phenotypes

Phenotypes are simulated based on a binomial draw from the logistic model in Eq 2.5. Given a set of causal SNPs with genotypes $\mathbf{X}_q$ and their effects $\boldsymbol{\beta}_q$, we first calculate the intercept necessary for an expected 50/50 ratio of cases to controls as $\mu = n^{-1}\mathbf{1}^{\mathrm{T}}(-\mathbf{X}_q^{\mathrm{T}}\boldsymbol{\beta}_q)$, calculate individual propensities $p_i = \text{logit}^{-1}(\mu + \mathbf{x}_q^{\mathrm{T}}\boldsymbol{\beta}_q)$, and then draw phenotypes as $Y_i \sim \text{Bin}(1, p_i)$.

Placing causal loci

To ensure a degree of confounding correlation between loci, we choose 5 causal SNPs at random but in a restricted manner from the LD blocks shown in Figure 2.1. Specifically, in each simulation trial, two SNPs are chosen from block 1 at random but subject to correlation $r \geq 0.4$, two SNPs are from block 2, also subject to $r \geq 0.4$, and one SNP is randomly chosen from block 3.

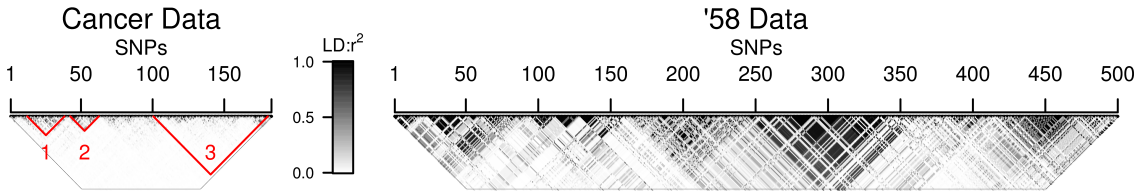

Figure 2.1: LD structure of the two genotype datasets used in the simulations. Shading indicates pairwise LD between SNPs, ranging from white ($r^2 = 0$) to black ($r^2 = 1$).

Simulation 1A: moderate effects

To aid an initial illustrative comparison between methods, our first study on the cancer data simulates a relatively constant effects structure. In each simulation trial we assign

a permutation of the effects (on the odds scale) $\exp\{\boldsymbol{\beta}_q\} = (1.287, 1.398, 1.246, 1.357, 1.419)$ to the selected five SNPs.

Simulation 1B: small effects

Providing a more challenging and variable set of causal targets, our second study on the cancer data randomly chooses causal SNPs as in 1A but draws each element β_{qj} of effects $\boldsymbol{\beta}_q$ independently as $\exp\{\beta_{qj}\} \sim \mathrm{N}(1.25(-1)^{\nu_j}, 0.02^2)$ with $\nu_j \sim \mathrm{Binom}(1, 0.5)$. The resulting effects are comparable to the small effects estimated in many GWAS studies (Manolio et al., 2009).

2.3.2 Simulation study 2: 1-7 loci in '58 data

The "'58" data is a complete-genotypes subset of data collected during the human GWAS for seven diseases described in WTCCC (2007). It comprises genotypes for 2199 subjects on 500 SNPs in the region 39.063723Mb-40.985321Mb on chromosome 22, this region being chosen by us as a contiguous run of markers that exhibits a mixture of high and low LD (Figure 2.1). To assess the how the number of causal SNPs affects the relative utility of modeling single versus multiple loci, we evaluated methods in seven distinct simulation substudies, simulating $1, \ldots, 7$ causal loci respectively. In each simulated trial of each substudy, the set of causal loci is chosen at entirely random from the 500 SNPs and the SNP effects are generated as in simulation 1B above.

2.3.3 Computation

Genotype imputation was performed using fastPHASE (Scheet and Stephens, 2006). All other analyses were performed in R (R Development Core Team, 2010), with the *glmnet* package (Friedman, Hastie and Tibshirani, 2010) used for fitting LASSO models. On a 2.4Ghz MacBook Pro with 4Gb RAM, on average 100 subsamples on the cancer data takes the following times: LLARRMA with permutation selection, 39.8s (sd=2.6s); LLARRMA with complement deviance selection, 389.2s (sd=64.6s); SS,

305.7s (sd=48.5s). Use of multiple/hard/dosage imputed data incurs negligible extra computation (assuming the imputation itself has been done in advance).

2.3.4 Competing methods

LLARRMA calculates a score (an RMIP) for each SNP in the study. We compare the ability of those scores to discriminate causal from non-causal SNPs with the SNP scores calculated by alternatives: the traditional GWAS approach of single locus regression, the LASSO-based subsample model averaging method stability selection (SS) recently proposed in a more general context by Meinshausen and Bühlmann (2010), a bayesian model averaging approach PIMASS (Guan and Stephens, 2011), and the standard approach of forward selection which is often used to better understand a hit region. As our approach and goal when analyzing GWAS data is significantly different from most methods, which greatly limits potential competing methods. Potential competitors tend to differ from our goal in one or more aspects, making them not appropriate for comparison within complex hit regions.

The first major difference is the complexity of the phenotype. Most methods are developed and tested on simpler phenotypes which would present with only a single causal SNP within a hit region (e.g., Zuber, Silva and Strimmer, 2012; Guan and Stephens, 2011; He and Lin, 2011; Motyer et al., 2011; Shi, Boerwinkle and Morrison, 2011; Alexander and Lange, 2011). As we assume a highly complex hit region with multiple correlated true signals, such methods would not be appropriate for this type of detailed analysis.

The second major difference between LLARRMA and its competitors is the scope of analysis. We focus only on a hit region, assuming that a method has first found a complex hit region, and try to dissect the hit region the best we can. Competitors take a genomewide approach and stress the ability to find hit regions rather than the actual underlying SNPs within them. This is typically observable in one of two ways.

The first, and most common, way that we are able to observe these differences is in how competitors define true and false positives. Specifically, it is common to define a true detection of the signal to be inclusion of any SNP within a region of the true signal, usually anything within some arbitrary physical distance or LD strength (see, for example, He and Lin, 2011; Guan and Stephens, 2011). A similar approach is often taken for false positives, where a set of false positives in close proximity are classified as a single false positive. Under many of these approaches, selection of any SNP in our simulations would be considered a true positive. The second way we can observe the strong genomewide focus is in the software, where in some cases it is required to have the full genome data to perform any analysis (e.g., GWASelect (He and Lin, 2011)).

Our method approaches characterizing model uncertainty when working within a frequentist framework, and applies it to a task for which it should be ideally matched. Frequentist literature typically ignores this problem while Bayesian literature explicitly models it. LLARRMA presents a solution for researchers who consider the problem crucially important but prefer not to address it within a Bayesian framework. In short, we condition on being frequentist, and examine the problem of variability or uncertainty of model choice working within that paradigm.

It is nonetheless interesting, however, to consider the performance of a contemporary Bayesian variable selection (BVS) method applied to the same simulations. See O'Hara and Sillanpää (2009) for a review of potential bayesian alternatives. We note that BVS methods attempt to answer a fundamentally different question, and so a comparison between with LLARRMA, Stability Selection or p-values from single locus regression, will not be strictly meaningful except in the most superficial sense. Specifically, Bayesian approaches model uncertainty about the parameters conditional on the data, whereas the frequentist approach models the expected variability of a statistic (such as an estimate from a model selection procedure) due to variation in different data

samples. Because the approaches ask different questions, even optimal implementations of each would not necessarily yield similar answers.

Single locus regression

We perform single locus regression with logistic regression as used in, for example, PLINK (Purcell et al., 2007). For each SNP, we fit a single-predictor version of Eq 2.5 and score its $-\log_{10} P$ ("logP"), where P is the p-value from a likelihood ratio test against an intercept-only model.

Stability selection

SS differs from LLARRMA in two main respects (see Figure 2.2). First, whereas LLARRMA selects variables within each subsample using a local (i.e., subsample-specific) penalty $\lambda^{(k)}$, SS uses a single global penalty λ applied to all K subsamples. Second, whereas LLARRMA chooses each $\lambda^{(k)}$ automatically, SS leaves its global λ as a free parameter. In SS, the RMIP (referred to as the "selection probability" in Meinshausen and Bühlmann, 2010) is thus left as a function of λ,

$$\widehat{\mathrm{RMIP}}_{\mathrm{SS}}(\lambda)_j = \frac{1}{K}\sum_{k=1}^{K} I(\hat{\beta}(\lambda; \mathcal{D}^{(k)})_j \neq 0) \tag{2.7}$$

giving rise to a sequence of RMIPs (a "stability path") for each locus j. Meinshausen and Bühlmann (2010) provide little guidance for choosing λ. As a choice is required to produce a unique RMIP and thereby ensure meaningful comparison with LLARRMA, we select λ to produce the stiffest possible competition: as the value that maximizes the criterion used for comparing methods. Specifically, given a criterion of success $u(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}})$ comparing truth $\boldsymbol{\gamma}$ with guess $\hat{\boldsymbol{\gamma}}$, we define

$$\hat{\lambda}_{\mathrm{oracle}} = \operatorname*{argmax}_{\lambda} u(\boldsymbol{\gamma}, \mathrm{RMIP}_{\mathrm{SS}}(\lambda))\,,$$

where "oracle" reflects the fact that choosing this unfairly advantageous value requires foreknowledge of $\boldsymbol{\gamma}$. We consider SS with and without the randomized LASSO, setting $\alpha = 0.2$ for the latter, and define the oracle by setting u to be the initial AUC (described below).

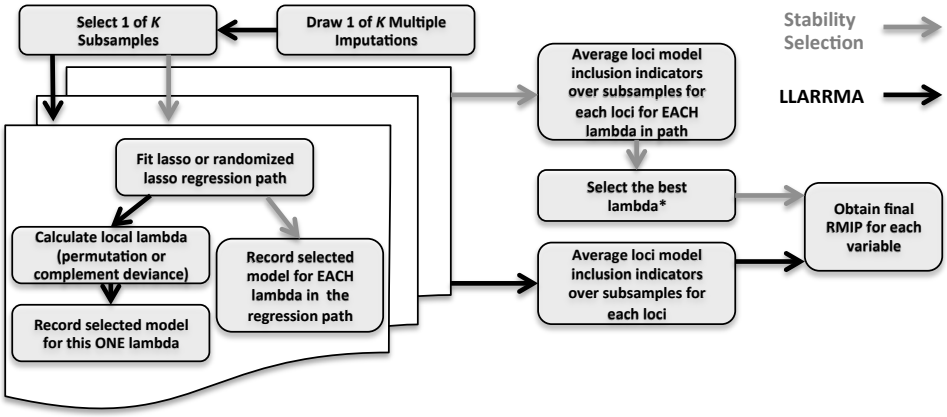

Figure 2.2: A comparison of LLARRMA and stability selection.

Forward Selection

Forward selection sequentially adds predictors into a model until a predefined stopping criterion is reached. Typically the criterion is based on a trade-off between the fit of the model and a some penalty on the number of predictors included (eg, AIC, BIC). A given penalty thus produces a single binary estimate of $\boldsymbol{\gamma}$ that would correspond to single point in a ROC plot. Providing a full ROC curve requires exploring a continuum of penalties of increasing stringency. However, unlike with the λ selection parameter in the LASSO, there is no obvious monotonic function that would accommodate the most popular stopping rules. Therefore, we use a simplistic stopping criterion that is monotonic: a bound on the number of included predictors, $q_{\max}$. Specifically, we identify a subset $q_{\max}$ of candidate SNPs associated with case-control status from m total SNPs using forward selection. Starting with a base model containing only an intercept term, we use binary logistic regression to test the significance of each SNP

in turn (as in single locus regression) to identify the most significant SNP. This SNP is then incorporated into the base model and the remaining markers are rescanned to identify the the most significant SNP conditional on the SNP(s) already in the base model. That process is repeated until the model contains $q_{\max}$ SNPs.

PIMASS

Guan and Stephens (2011) describe a Bayesian variable selection method suitable for QTL identification in genomewide association studies. The method, an Markov chain Monte Carlo sampler implemented in their computer program PIMASS[1], uses a hierarchical prior to model jointly the number of included SNPs and the sizes of their effects. We run PIMASS based on a logit link (option `-cc`) under standard settings, using a 100,000 sample burnin[2] (`-w 100000`) and obtaining results based on 1,000,000 samples (`-s 1000000`). For each SNP, PIMASS provides a Bayes factor (BF) quantifying support for its inclusion in the model.

Other methods

We considered including BIMBAM (Servin and Stephens, 2007), which not only models multiple loci but also imputation uncertainty. However, we found that applying it to our smallest dataset incurred a severe computational overhead that precluded fair comparison (eg, a single run on the cancer data that allowed up to 3 SNPs to enter the model required 34 hours).

---

[1] http://www.bcm.edu/cnrc/mcmcmc/pimass

[2] Experimenting with a longer burnin of 1,000,000 gave similar results, suggesting that 100,000 is adequate.

2.4 Simulation Results

2.4.1 Simulation study 1A: moderate LD, moderate effects

We simulated 1000 case-control datasets based on the cancer data (see Methods and Figure 2.1). Each simulated dataset had approximately balanced cases and controls, with individuals' outcomes influenced by 5 SNPs of moderate effect (odds ratios 1.246-1.419) out of 183 SNPs in total, and existed in both a complete form, referred to as the "complete" data set, and an incomplete form, in which some genotype values were set to be missing. The incomplete form was available in three alternative imputations: a "hard" imputation, a "dosage" imputation, and an ensemble of 100 sampled imputations that constituted a single "multiple" imputation set (these imputations being generated by fastPHASE, Scheet and Stephens, 2006). At each simulation we tested different analysis methods that each produced a score per SNP. Our subsequent comparisons of those methods were based on how well their scores discriminated the 5 SNPs that were causal from the 178 that were not. The seven method examined were (*short names in parentheses*): single SNP logistic regression (*single locus regression*); LLARRMA using permutation selection and ordinary LASSO (*permutation selection*); LLARRMA using complement deviance selection and ordinary LASSO (*complement deviance*); stability selection using ordinary LASSO and oracle penalization (*oracle stability selection*); PIMASS (*pimass*); conditional regression scans (*forward selection*). All methods were applied to the complete data, and most (all but PIMASS and forward selection) were applied to the hard imputation, and dosage imputation versions of each simulated data set; resample-based methods (ie, LLARRMA and SS), which were set to use $K = 100$ subsamples, were also applied to the multiple imputation set.

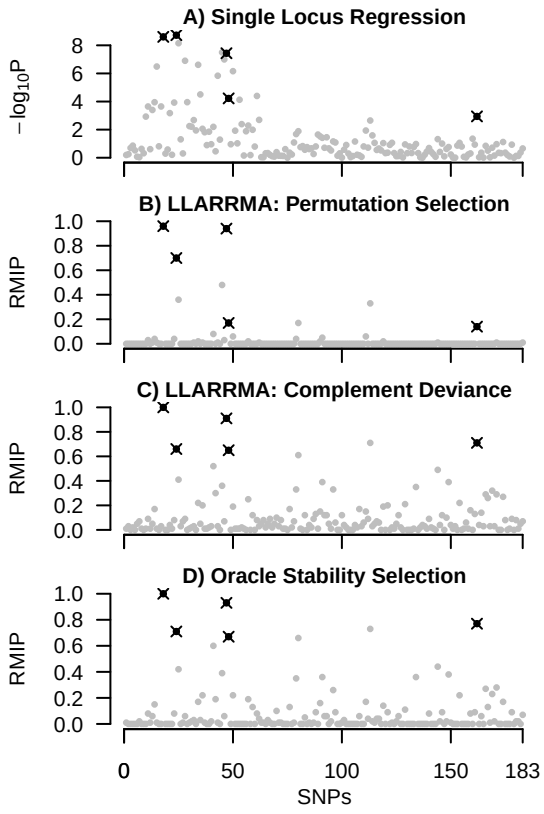

Figure 2.3: Results for seven procedures applied to an example case-control data set from simulation study 1A. Plots show SNP score (logP or RMIP) against SNP location in the cancer data, with causal SNPs in black and non-causal SNPs in gray.

An example simulation

Figure 2.3 plots SNP location against SNP-score for each method in an example simulation applied to complete data. Causal SNPs are plotted as black crosses and non-causal SNPs as gray dots. In single locus regression (Figure 2.3a), SNPs are scored as $-\log_{10} P$ (logP; see Methods). Although the causal SNPs between 1-50 tend to attract higher scores, so do many of the non-causal SNPs between 1-60, giving rise to a cloud of association that is characteristic of many hit regions in real GWASs. The remaining methods (2.3b-g) report inclusion probabilities (RMIPs) for each SNP. These describe a frequentist probability that each SNP would be included in a sparse model that seeks to estimate the joint effects of multiple SNPs. Because SNPs compete with each other

for inclusion in these methods, the resulting scores more clearly differentiate the SNPs. In this example, that increased sparsity coincides with the set of higher scored loci being more enriched for causal SNPs than is the case with single locus regression.

Results from 1000 simulations

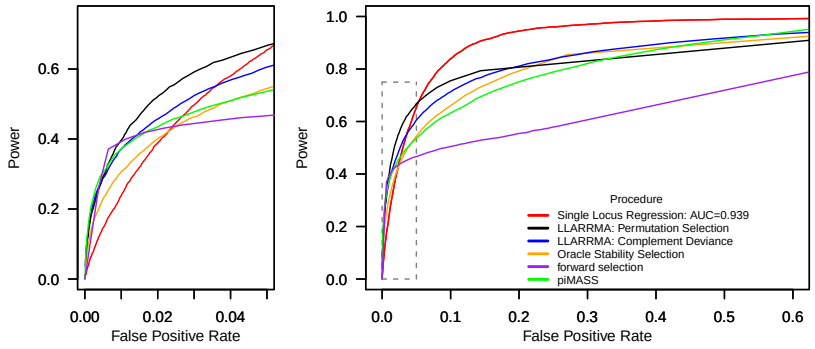

Figure 2.4: ROC curves for simulation study 1A: moderate SNP effects in a hit region of moderate LD. Curves compare the ability of seven methods to discriminate causal from non-causal loci in 1000 simulated case-control data sets. Right plot shows the full ROC curve; left plot shows a zoomed section focusing on the top-scoring SNPs of each method.

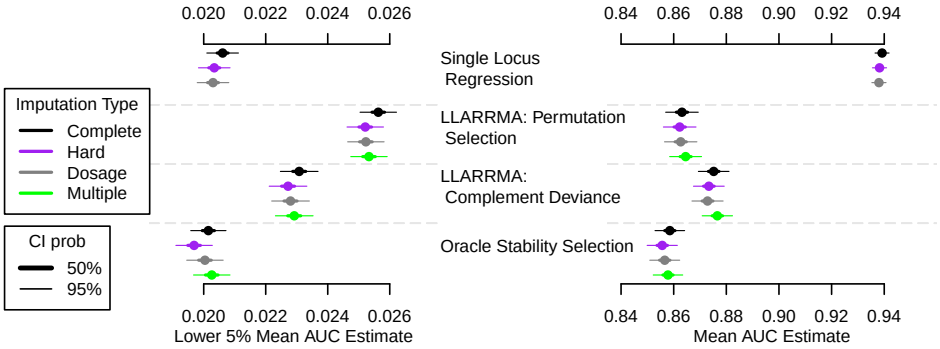

Figure 2.5: Area under the ROC curve (AUC) for seven methods applied to four types of imputed genotype data in simulation study 1A: moderate SNP effects in a hit region of moderate LD. Each AUC estimate is based on 1000 simulations and is plotted as mean (dot), 50% CI and 95% CI.

Figure 2.4 plots ROC curves (see Methods) for each of the methods, with single locus regression applied to complete genotype data and resample-based methods applied to genotype data with ∼10% missingness that has been multiply imputed (see Methods

and below). The ROC curve plots the trade-off between power (the proportion of causal SNPs declared as influential) and false positive rate (FPR; the proportion of non-causal SNPs declared as influential) when thresholding the SNP scores (logPs or RMIPs) at different values. The initial ROC is arguably of greater relevance to GWA studies than the full ROC because it focuses on enrichment of causal signals among the top-scoring SNPs. A method whose top four SNPs are causal, but which never finds the fifth causal SNP, is arguably more valuable than one whose top SNPs are non-causal but which finds all five causal loci among its middle scoring SNPs. Figure 2.4 shows both the full ROC curve (right) and the initial ROC curve (ie, where FPR $\leq 5\%$; left). Figure 2.5 plots the area under the curve (AUC) for the initial and full ROC curves for all seven methods under four conditions: where the available genotype data is complete, or has $\sim$10% of its genotypes missing but available in hard-, dosage- or multiply- imputed forms. All point estimates (plotted curves in Figure 2.4 and mean AUCs in Figure 2.5) are based on averages over the 1000 simulations.

Figure 2.4 shows that single locus regression most powerfully discriminates causal from non-causal SNPs when the experimenter is prepared to follow up to 10% or more of the available SNPs, but in scenarios where at most the top 5% of SNPs would be considered it is dominated by LLARRMA's permutation selection, and for more restricted sets it is dominated by complement deviance and oracle SS. We also observe that pimass and forward selection perform very well in the first third of of the initial ROC but their performance quickly drops off when compared to LLARRMA. Figure 2.5 echoes these trends. It also shows how the methods perform under different forms of imputation, although no consistent pattern emerges favoring one form over the others.

2.4.2 Simulation study 1B: moderate LD, small effects

We performed a second set of simulations with a design identical to 1A above except with smaller SNP effects (odds ratios around 1.25). The results in Figures 2.6 and 2.7

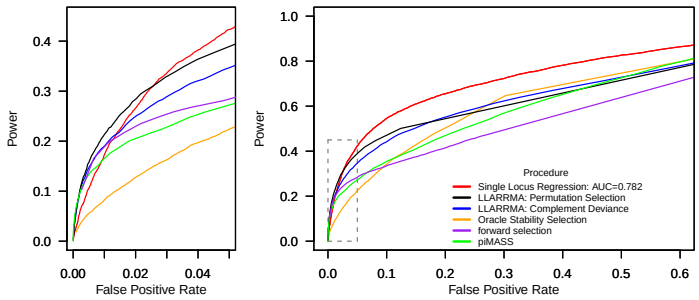

Figure 2.6: ROC curves for simulation study 1B: small SNP effects in a hit region of moderate LD. Curves compare the ability of the methods to discriminate causal from non-causal loci in 1000 simulated case-control data sets. Right plot shows the full ROC curve; left plot shows a zoomed section focusing on the top-scoring SNPs of each method.

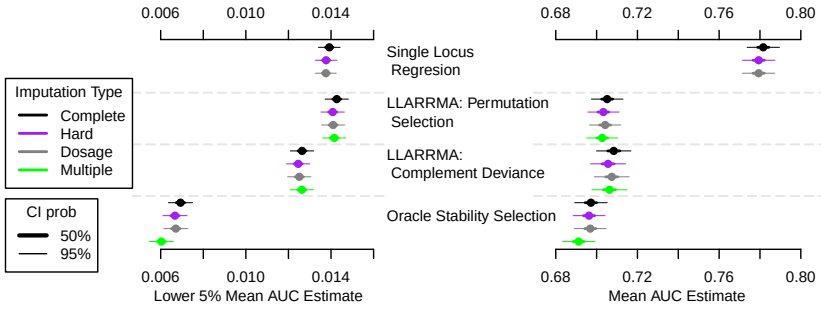

Figure 2.7: Area under the ROC curve (AUC) for the methods applied to four types of imputed genotype data in simulation study 1A: moderate SNP effects in a hit region of moderate LD. Each AUC estimate is based on 1000 simulations and is plotted as mean (dot), 50% CI (thick bar) and 95% CI (thin bar).

show that although some of the LLARRMA methods dominate in the first third of the initial ROC curve, they generally offer little improvement over single locus regression under these conditions. The poor performance of the oracle SS method is striking. "Oracle" refers to the fact that an aspect of how these methods were applied required foreknowledge of the answer: namely, their free parameter λ was set to maximize their initial AUC. That all of the LLARRMA variants, none of which have the "oracle" advantage, dominate both oracle SS methods suggests a systematic shortcoming of stability selection in this setting. Figure 2.8 helps explain the phenomenon. For each of a representative subset of the 1000 simulations, it plots the value of λ chosen by

oracle SS (black plus). This λ is then applied to the LASSO paths of all $K = 100$ subsamples, and hence is "global", to give the final RMIP scores per SNP. For the same simulations and subsamples, Figure 2.8 also plots the $\lambda^{(k)}$ chosen by LLARRMA "locally" for each subsample $k = 1, \ldots, K = 100$ (gray crosses), with this choice based on the complement deviance criterion, which maximizes out-of-sample predictions. If the $\lambda^{(k)}$ were truly optimal for the LASSO fit to each subsample, then this illustrates how even the best choice of a global λ would translate to a suboptimal local $\lambda^{(k)}$ for most subsamples.

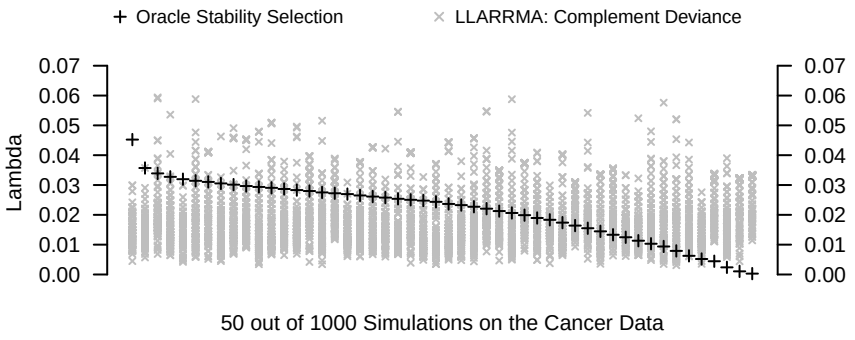

Figure 2.8: Global choice of penalty parameter λ by oracle stability selection (black pluses) versus local, per-subsample, choice by LLARRMA complement deviance selection (gray crosses) in 50 representative simulation trials out of 1000 performed for simulation study 1B.

2.4.3 Simulation study 2: strong LD, small effects

To examine the relative performance of the single and multiple locus methods in a more challenging setting, we simulated 700 case-control data sets based on the '58 data, a region on chromosome 18 containing blocks of strong LD from the GWAS of WTCCC (2007) (see Methods and Figure 2.1). Each simulated dataset had a complete set of genotypes and approximately balanced cases and controls. Individual's outcomes were influenced by between 1 and 7 causal SNPs of small effect (odds ratios around 1.25), with 100 simulations devoted to each simulated number of causal loci $m_q = 1, \ldots, 7$.

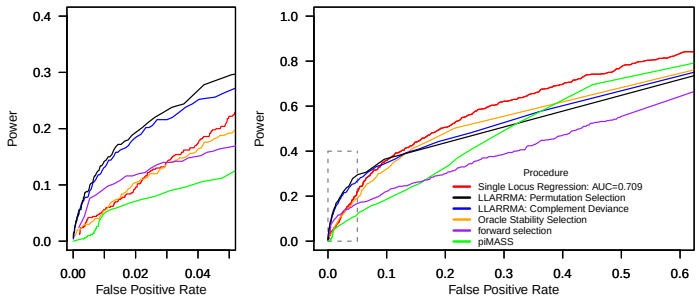

Figure 2.9: ROC curves for simulation study 2 with 5 loci: small SNP effects in a hit region of strong LD. Curves compare the ability of seven methods to discriminate causal from non-causal loci in 100 simulated case-control data sets. Right plot shows the full ROC curve; left plot shows a zoomed section focusing on the top-scoring SNPs of each method.

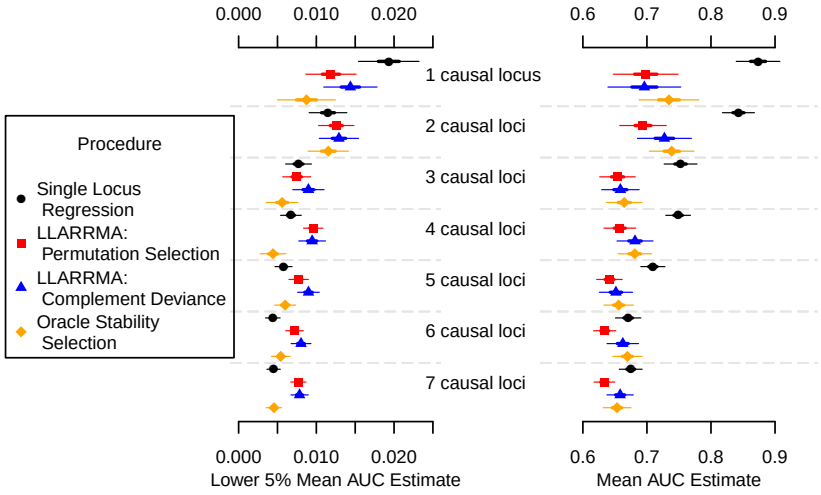

Figure 2.10: Area under the ROC curve (AUC) for 7 methods applied to simulated case-control influenced by 1-7 causal loci in simulation study 2: small SNP effects in a hit region of strong LD. Each AUC estimate is based on 1000 simulations and is plotted as mean (dot), 50% CI and 95% CI.

Figure 2.9 shows the initial and full ROC curves from the 100 simulations in which 5 causal loci were simulated. In this high correlation - low signal setting, all forms of LLARRMA dominate single locus regression in the initial ROC curve, suggesting an advantage of simultaneously modeling multiple loci in the presence of high LD. By contrast, both forms of oracle SS equal or underperform single locus regression, a result similar to that in 1B above, suggesting that this modeling is suboptimal in

SS. Figure 2.10 summarizes results from all 700 simulation trials and shows the effect of varying the number of causal loci. With one causal SNP, single locus regression equals or betters any other method; but as the number of causal SNPs increases, its advantage over multiple locus methods diminishes. In particular, for four or more loci the standard LLARRMA methods (permutation selection and complement deviance) consistently outperform in the initial AUC, whereas the oracle SS methods consistently underperforms both LLARRMA and single locus regression.

2.5 Discussion

We present a general approach for characterizing frequentist variability in LASSO-based model choice, LLARRMA, and apply it to a problem for which it should be well suited: discriminating true from false signals among a set of SNP predictors that are often highly correlated due to LD. In doing so, we evaluate two criteria for automatically choosing the LASSO penalization parameter λ (permutation and complement deviance selection), demonstrate potential superiority of local vs global regularization in subagging (through comparison with stability selection) and propose a natural way to combine resampling aggregation with multiple imputation to account more comprehensively for different sources of variability in model choice.

LLARRMA's intended use is in focused analyses on hit regions that have been already identified during whole genome analysis. Rather than replacing single locus regression, its value lies in what it subsequently adds to that analysis. When there are few causal SNPs and mild LD, the best LLARRMA methods add little. However, as shown for the '58 data (simulations 2B), when there the causal SNPs are many (≥ 4 in our simulations), applying LLARRMA produces a top set of loci that are enriched for causal signals relative to logPs from single SNP association. Of our two alternatives for automatically selecting the penalty λ, we found a slight but consistent advantage of

permutation selection (modified from Ayers and Cordell, 2010). This could reflect its discovery-based motivation matching our discovery-oriented evaluation, and does not preclude complement deviance being superior in predictive settings.

We explored the use of stability selection (SS; Meinshausen and Bühlmann, 2010) in this context but find it no better than, and usually inferior to, LLARRMA, despite the fact that our evaluation of SS is based on an optimal calibration of its (unspecified) penalization parameter. One explanation is that SS's use of a single global λ for all subsamples *underfits* the data, in that it fails to accommodate structural differences between LASSO paths fit to different subsamples. The local automatic regularization in LLARRMA implies a different perspective: that λ is a parameter intrinsic to, and only meaningful in the context of, a single LASSO path on a single (subsampled) realization of the data. Another factor could be our evaluation scheme: by calculating power and FPR at different thresholds of RMIP, we (reasonably, in our view) assume that RMIPs should be comparable across simulated data sets. However, when we threshold instead on the ranks of the RMIPs within simulated data sets (such that the best RMIP in trial $s = 1$ is equivalent to the best RMIP in $s = 2$), the performance gap between SS and LLARRMA narrows (data not shown), suggesting SS RMIPs are discriminatory but their absolute values are less comparable across studies. Lastly, although our implementation of SS uses subsample proportion $\phi = 2/3$ rather than the original $\phi = 1/2$ of Meinshausen and Bühlmann (2010), our preliminary studies (not shown) do not suggest this biases comparisons with LLARRMA.

Alexander and Lange (2011) recently demonstrated SS's inferiority to single locus regression for identifying unlinked QTLs in whole genome association (also using data from WTCCC, 2007). The weakness we identify in SS may help explain that poor performance. Nonetheless, we believe that to expect SS (or LLARRMA for that matter) to beat SLR at its own game is not only optimistic, especially given the near-optimality

of marginal approaches suggested by Fan and Lv (2008), but also distracts from the potential advantages of multi-predictor shrinkage for disentangling highly correlated signals in LD blocks following an initial SLR scan.

Multiple imputation is simply accommodated by our resampling scheme, with draws from an arbitrarily complex imputation algorithm dovetailing naturally with the drawing of each subsample from the full data. However, our results suggest that even with 10% missing genotypes multiple locus inference is served just as well by simpler "plug-in" imputation estimates (hard and dosage). Nonetheless, we advocate multiple imputation where possible because it more comprehensively models imputation uncertainty (among genotypes or other covariates) that could be more pronounced in messier data sets.

Resample aggregation techniques such as bootstrap aggregation ("bagging"; Breiman, 1996) or subsample aggregation (subagging; Bühlmann and Yu, 2002) have been found to produce estimates of $\boldsymbol{\gamma}$ that are more stable than from a single estimation run in the sense that those estimates have lower frequentist risk under squared error loss (Bühlmann and Yu, 2002). However, we prefer subagging (as in Valdar et al., 2009; Meinshausen and Bühlmann, 2010) for two reasons. First, theoretical results in Politis, Romano and Wolf (1999, p. 47-51) suggest that subsampling is less efficient but more general than bootstrapping; specifically, that whereas bootstrap methods must often assume that the estimated statistic is at least locally smooth (which the true or sampled $\boldsymbol{\gamma}$ is not), this assumption is not needed for subsampling. Second, resampling individuals with replacement (bootstrapping) poorly approximates variation in GWAS samples because whereas bootstrapping produces frequent duplicates, observing multiple individuals with identical genetic composition is typically highly improbable.

Although we describe LLARRMA in the case-control setting using the logistic model, it is easily extended to the analysis of quantitative traits or any response to

which the LASSO can be applied. Similarly, although we model under the simplistic assumption of additive effects and no local epistasis, these assumptions could be relaxed by a more sophisticated specification of locus effects, for example, using the group LASSO (Yuan and Lin, 2006*b*; Meier, Geer and Bühlmann, 2008) or a similar structured penalization scheme.

In summary we describe an approach for characterizing frequentist variability of model choice in binary data that can be usefully applied to the reprioritization of SNPs in hit regions of a case-control GWAS. The method uses LASSO local automatic regularization resample model averaging (LLARRMA) and integrates well with schemes for imputation of missing data. We provided an implementation of LLARRMA in an R-package R/llarrma.

Chapter 3

Generalization of Resample Model Averaging

In this chapter we will discuss the motivation for generalizing the resample model averaging (RMA) framework used in LLARRMA (see Chapter 2). Specifically we will focus on the short comings of the assumed model under which LLARRMA was designed. In addressing this issue, we will expose other issues related to RMA, and present a generalized RMA framework which addresses the underlying problems.

3.1 Introduction

Analyses of human genome wide association studies (GWAS) have concentrated on modeling effects of single loci. Multiple locus alternatives do exist, and they can produce more robust and powerful results (Wang et al., 2012). Unfortunately, multiple locus methods are seldom used in practice. These approaches include multiple locus regression by forward selection (Cordell and Clayton, 2002), Bayesian model selection (Stephens and Balding, 2009), penalized regression approaches (Malo, Libiger and Schork, 2008; Cule, Vineis and De Iorio, 2011), and resampling based methods (Valdar et al., 2012; Alexander and Lange, 2011; Guy, Santago and Langefeld, 2012).

The assumed underlying genetic model can greatly determine the success of GWAS methods. The statistical modeling of GWAS for complex traits most often assumes that

the effect of the minor allele is strictly additive with respect to its count. Although additive models are often adequate for modeling, the underling phenotype architecture may not be additive. Two such underling architectures are dominant alleles and heterosis (or overdominance) effects. When the minor allele is dominant, additive models do not lose much power, but when the major allele is dominant (recessive traits) there is a great loss in power when considering additive models (Wang et al., 2012; Kim et al., 2010). In diseases such as Cystic Fibrosis and Phenylketonuria, the heterozygote effect is indistinguishable from that of the major allele homozygote, i.e., the major allele is dominant with respect to the minor allele. Many other Mendelian diseases follow this pattern of a dominant major allele. Heterosis, or an advantage to the heterozygote, can be found in plants and model organisms (Neale et al., 2008). Although complex traits have a more complicated genetic architecture than Mendelian diseases, it is reasonable to expect that some of the loci involved may follow similar dominance or heterosis effects. Methods under which the assumed genetic model allows for such deviations from additivity can be more powerful in detecting underline causal variants. Whereas it may be easy to model more complex models such as dominance for single locus methods (e.g. Servin and Stephens, 2007; Yeager et al., 2007; Li et al., 2009), multiple locus modeling can be less straightforward.

When considering these non additive effects, the sample sizes needed to detect the deviations from additivity are rather important. Recently, GWAS have incorporated substantially larger sample sizes, which may lead to the ability to better model non-additive effects. Specifically, due to the nature of genotype probabilities, much larger samples can be required to detect a dominant minor allele, as the homozygous minor allele (the only genotype with signal) is a rare genotype for even moderate minor allele frequencies. While dominant major alleles also require larger sample sizes to detect

their deviations from additivity, an additive model is more likely to detect their signals than the signals of a dominant minor allele.

In Chapter 2 we introduced a resample model averaging (RMA) based method called LLARRMA for the analysis of human GWAS hit regions. LLARRMA characterizes sensitivity of locus choice due to sampling variability and provides LASSO shrinkage that is automatically regularized through either a predictive- or discovery-based criteria. We also introduced a way to use multiple imputation within the resampling to account for imputation uncertainty, which adds very little computation complexity to the method. We showed that the reprioritization given by LLARRMA (using either selection criteria for the LASSO penalty parameter) enriched the top set of loci for true signals when compared to single locus regressions. We also examined the use of stability selection (SS; Meinshausen and Bühlmann, 2010) for GWAS hit regions. We found that LLARRMA dominated the performance of SS; results that are consistent with Alexander and Lange (2011), who have recently proposed SS for whole GWAS data rather than hit regions.

The statistical modeling of dominance can introduce complexities that are comparable to modeling of rare variants and haplotypes. Specifically, in order to model dominance, more than one predictor must be included in the model for each locus and are often highly correlated. This shares some of the complexities of modeling haplotypes. That is, when modeling haplotypes there will be multiple predictors to model each haplotype region, and these predictors are often collinear. Furthermore, when we examine the modeling of dominance, the presence of the homozygous minor allele genotype is essential for distinguishing additive from dominant effects. As this genotype is present at the rate of the square of the minor allele frequency (which quickly becomes rare for even common variants), its easy to see that dominance predictors present with similar problems rare variants.

In this chapter, we propose a generalization of the LLARRMA method of Valdar et al. (2012), for reprioritizing genetic associations in a hit region of a human GWAS. We describe a principled extension that allows modeling of non-additive effects such as dominance. In doing so, we identify two important problems. The first is that when considering rare predictors (e.g. the dominance predictor or rare variants) with subsampling based RMA, predictors may become monomorphic, reducing the dimensionality of the data on the subsample. The second issue is that the LASSO penalty does not properly utilize the additional predictors introduced for modeling more general effects such as dominance. To address these problems, we propose a modified resampling procedure based on continuous weights for the subjects, which eliminates the issue from subsampling. Further, we introduce a group penalty thats combines the multiple predictors that represent a single locus to better unitize the available information for each locus. We show that when multiple correlated SNPs are present in a hit region (identified by for example, standard single locus regression) our generalization produces a reprioritization that is enriched for true signals.

3.2 Methods

3.2.1 Assumptions and Statistical Model

We start by considering the use of standard linear regression to estimate the effects of m SNPs (in a hit region) on a quantitative outcome from n individuals. We then describe statistical approaches to identify a subset of m_q SNPs that might be truly influential. Here we define a "true signal" to be a SNP that most strongly tags an underlying causal variant, a "background" SNP to be a SNP that is not a true signal, and an optimal analysis as one that distinguishes true signals from background SNPs within a hit region. We assume that the hit region has been previously identified by an initial genomewide scan (using, for example, single locus regression), that the m

SNPs may be in high LD, and that $m_q < m < n$. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be an n-vector of quantitative responses. Let $\mathbf{X}$ to be the $n \times m$ matrix of unphased SNP genotypes {qq,qQ,QQ} where Q is the minor allele. To define our generalized model which can model both additive and dominance effects simultaneously, we will consider two data matrices that are functions of $\mathbf{X}$. Specifically, we define the $n \times m$ matrix $\mathbf{A}$ by the count of the minor allele, i.e. SNP genotypes coded as $\{0, 1, 2\}$ for unphased genotypes {qq,qQ,QQ}. $\mathbf{A}$ is the matrix that incorporates the additive portion of the effects. We also define the $n \times m$ matrix $\mathbf{D}$ where $d_{ij} = I\{X_{ij} = qQ\}$, which incorporates a deviation from additivity which can detect dominance effects. We note that our main goal is not to distinguish which loci are additive or dominant, but rather simply identify which loci are true signals. We then let $\mathcal{D} = \{\mathbf{y}, \mathbf{A}, \mathbf{D}\}$ be the data considered for modeling. Let $\mathcal{N} = \{1, \ldots, n\}$.

We model the quantitative phenotype of individual i by a linear regression of the $2m$ predictors as

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{A}^T \boldsymbol{\beta}_a + \mathbf{D}^T \boldsymbol{\beta}_d + \epsilon \,, \tag{3.1}$$

where μ is the intercept, $\boldsymbol{\beta}_a, \boldsymbol{\beta}_d$ are the effects of the m predictors corresponding to data matrices $\mathbf{A}, \mathbf{D}$ respectively, and $\epsilon_i \sim N(0, \sigma)$ are Gaussian errors.

In a likelihood based model, one may prefer an equivalent model in which each genotype is assigned its own effect, i.e. an ANOVA model. When considering a penalized regression framework, the performance of our proposed model (Eq 3.1) and a penalized ANOVA can differ. Our tests suggest that our model outperforms the ANOVA model.

We assume that only a subset of the m SNPs have a genuine effect on the response, and define the corresponding vector of 0-1 inclusions $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ such that $\gamma_j = I(\beta_{a,j} \neq 0 \text{ or } \beta_{d,j} \neq 0)$ which identifies loci having a genuine effect. Variable selection methods are common ways to infer $\boldsymbol{\gamma}$, and to thereby also estimate the identity of the true signals. The hard estimate $\hat{\boldsymbol{\gamma}}$ obtained form a single variable selection, although

potentially consistent, is a statistic of high variance as it fails to capture information about how sensitive the selections are to sampling variability (Valdar et al., 2012).

We seek to estimate $\boldsymbol{\gamma}$ in way that incorporates uncertainty in model choice arising through, for example, potential variability of the selected set due to finite sampling. One approach for doing this is resample model averaging (RMA; Valdar et al., 2009, 2012). In RMA, one applies a model selection procedure to repeated resamples of the data, to simulate potential differences observed in a different sampling, and base subsequent inference on the aggregate of these results. In Chapter 2 we use a subsampling-based RMA. Although it is an attractive approach, it can be problematic when dealing with more rare predictors, such as **D** here. The presence of the QQ genotype is necessary to differentiate between **A** and **D**. With subsampling based RMA, on a given subsample a locus may lose any subjects observing QQ, changing the dimensionality of the data. On such subsamples, we are unable to test for dominance effects at such loci. For loci with a low enough MAF, one may also lose the ability to infer anything about the locus if the locus becomes monomorphic on the subsample. To avoid these problems, we propose a generalization that retains every individual's information at some level in the resampled model's fitting and approaches subsampling in the extreme case.

We considered the use of LASSO penalized regression (Tibshirani, 1996) for variable selection, as done in LLARRMA (Valdar et al., 2012). Due to the nature of their definitions, the **A** and **D** predictors are always highly correlated with each other. This is problematic as the LASSO tends to select only one of the predictors for a given locus. In order to better incorporate the information that is poorly utilized by the LASSO, we propose to use the group LASSO (Yuan and Lin, 2006*a*).

3.2.2 Generalized resample model averaging

Rather than obtaining a binary estimate of each γ_j, we instead seek to estimate its expectation $\mathrm{E}(\gamma_j)$ over weighted resamples, hoping to approximate its expectation over

samples from the population. We start by drawing K random weights $\mathbf{w}_1, \ldots, \mathbf{w}_K$ where $\mathbf{w}_k = (w_{1_k}, \ldots, w_{n_k})$ with $w_{i_k} \overset{iid}{\sim}$ Weighting$(\cdot)$ to be dscussed later. Each weighted resample comprises data $\mathcal{D}^{(k)} = \{\mathbf{y}, \mathbf{A}, \mathbf{D}, \mathbf{w}_k\}$. For each weighted resample k, we apply a model selection procedure to produce $\hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)}) = \hat{\boldsymbol{\gamma}}^{(k)}$, the m-length binary vector of estimated loci inclusions based on the kth weighted resample. Applying this to K weighted resamples gives the $m \times K$ matrix $\boldsymbol{\Gamma} = [\hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\gamma}}^{(2)}, \cdots, \hat{\boldsymbol{\gamma}}^{(K)}]$. The proportion of times that the jth predictor is included in the resample based models is given by

$$\widehat{\text{RMIP}}_j = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\gamma}}(\mathcal{D}^{(k)})_j = \frac{1}{K}\sum_{k=1}^{K} \hat{\gamma}_j^{(k)} = \frac{1}{K}\sum_{k=1}^{K} \Gamma_{jk}\,, \tag{3.2}$$

which we refer to as its resample model inclusion probability (RMIP).

Generalized RMA weights

We propose to model $w_{i,k} \overset{iid}{\sim}$ Weighting$(\cdot)$ where Weighting$(\cdot) = \text{U}(0, 1)$. By doing so, we obtain continuous observation weights that ensure that each observation carries some weight in the model fitting, ensuring that the dimension of the data remains constant on each resample. Under this setting, $E(\mathbf{1}^{\text{T}}\mathbf{w}_k) = \frac{1}{2}n$, indicating that on average, each resample uses half of the available data. One may consider the interpretation of $w_{i,k}$ as the proportion of observation i used in resample k.

To connect the weights, $\mathbf{w}_k$, of the generalized RMA to the subsampling based RMA framework of Valdar et al. (2012), we consider $\mathbf{w}_k \overset{iid}{\sim}$ Subsampling(ϕ), where Subsampling(ϕ) is a function that returns a n-vector of 0's and 1's with ϕn 1's. Here, we consider $\phi = 1/2$ so that the amount of data used on each resample is consistent with our proposed weighting method.

Our generalization of the discrete subsampling weights to continuous uniform weights is similar in spirit to how bootstrap weights were generalized to continuous weights in the Bayesian bootstrap (Rubin, 1981). Specifically, one can represent the bootstrap

in the generalized RMA setting by using $\mathbf{w}_k \overset{iid}{\sim}$ Bootstrap$(\cdot)$, where Bootstrap$(\cdot)$ is a draw from a multinomial distribution with n groups, each equally probable. Under the bootstrap weights, $w_{i,k} \in \{0, 1, \ldots, n\}$ with $\mathbf{1}^{\mathrm{T}}\mathbf{w}_k = n$. As some weights will be zero on a given resample, the bootstrap has the same potential issue the data's dimension may be reduced when excluding subjects that presented under subsampling. The Bayesian bootstrap, a generalization the bootstrap, proposed to model $\mathbf{w}_k \overset{iid}{\sim}$ Dirichlet(1), where Dirichlet(1) is a uniform dirichlet distribution. The Bayesian bootstrap obtains similar results to the bootstrap, but would avoids the issues arising from weights of 0 in the bootstrap.

Selection within a resample using the group-LASSO

The group-LASSO estimates $\boldsymbol{\beta}$ for subsample k via the minimization

$$\hat{\boldsymbol{\beta}}^{\mathrm{grp}}(\lambda; \mathcal{D}^{(k)}) = \underset{\boldsymbol{\beta}_a, \boldsymbol{\beta}_d}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\beta}_a, \boldsymbol{\beta}_d; \mathcal{D}^{(k)}) + \lambda \sum_{j=1}^{m} \sqrt{{\beta_{a,j}}^2 + {\beta_{d,j}}^2} \right\}, \tag{3.3}$$

where ${\hat{\boldsymbol{\beta}}^{\mathrm{grp}}}^{\mathrm{T}} = [{\boldsymbol{\beta}_a}^{\mathrm{T}}, {\boldsymbol{\beta}_d}^{\mathrm{T}}]$, λ is a penalty parameter, and $\ell(\boldsymbol{\beta}_a, \boldsymbol{\beta}_d; \mathcal{D}^{(k)})$ is the weighted log-likelihood of $\boldsymbol{\beta}_a$, $\boldsymbol{\beta}_d$, and data $\mathcal{D}^{(k)}$ given by,

$$\ell(\boldsymbol{\beta}_a, \boldsymbol{\beta}_d; \mathcal{D}^{(k)}) = \sum_{i=1}^{n} w_{i_k} log(f(\beta_a, \beta_d; y_i, \mathbf{x}_i))$$

where $f(\beta_a, \beta_d; y_i, \mathbf{x}_i)$ is the Gaussian likelihood. We note that the Gaussian log likelihood may be replaced by the sum of squares for quantitative phenotypes, but we have presented the material in the likelihood form to show the generality of the procedure. Also, we note that when a group consists of a single predictor the group LASSO penalty simplifies to the LASSO penalty, i.e. $\sqrt{\beta^2} = |\beta|$. Although our model can account for various types of effects, as the group LASSO estimate requires that each member of the group has either a zero or a nonzero coefficient, we are unable to distinguish which

loci have additive or dominant effects. However, this is not an issue as our main goal is to identify true loci, not their underling effect types. To arrive a single estimate of $\boldsymbol{\gamma}$, as required for model averaging, we must devise a suitable criterion for choosing the penalty λ. We propose to identify a value $\lambda^{(k)}$ specific to weighted resample k (ie, local) by permutation selection.

Discovery-based selection of $\lambda^{(k)}$: permutation selection

We continue to use the permutation selection criterion described in Chapter 2. Specifically, given a weighted resample k, we estimate for a given permutation of the response, $\boldsymbol{\pi}(\mathbf{y}^{(k)})$, the smallest penalty, $\lambda_0(\boldsymbol{\pi}, k)$, required to zero out all predictors. We note that observation weights (and randomization of penalty, see next section) are fixed under permutations. For each of S permutations $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S$, the permutation selection λ for weighted resample k is defined to be

$$\hat{\lambda}^{(k)} = \text{median}(\lambda_0(\boldsymbol{\pi}_1, k), \lambda_0(\boldsymbol{\pi}_2, k), \ldots, \lambda_0(\boldsymbol{\pi}_S, k)) \,. \tag{3.4}$$

Model selection among highly correlated SNPs: the randomized group-LASSO

A new generalization of the LASSO, the randomized LASSO, was presented in Meinshausen and Bühlmann (2010). Whereas the LASSO penalizes the absolute value of the coefficients proportional to the penalty λ, the randomized LASSO changes the penalty of each coefficient to a random value in $[\lambda, \lambda/\alpha]$. The weights of the penalty are reminiscent of the adaptive LASSO (Zou, 2006), but with the perturbation being random rather than based on previous estimates. Applying the randomized LASSO many times (e.g. when used in stability selection or RMA) and considering variables which are often chosen can be rather powerful.

We have found that the randomized LASSO can be quite useful in the RMA framework, especially among highly correlated data. The LASSO can produce unstable

estimates when in the presence of highly correlated data, and minor changes in the data may cause the LASSO to switch included predictors. Although the resampling of RMA allows us to see how the LASSO performs over a number of slightly different data realizations, these realizations are conditioned upon the observed sampling (i.e. within sample variability). When considering highly correlated SNPs, their may only be a few subjects for which the SNP values differ. When considering such SNPs, it may not be clear which SNP is clearly the true signal, i.e. an independent second sample may switch the preferred SNP. The additional perturbations of the randomized LASSO may allow for the preferred SNP by the LASSO to switch within the observed resampling. The use of such randomized penalty comes with a tuning parameter which may need calibration specific to the data.

We follow the motivation of Meinshausen and Bühlmann (2010) and propose a new generalization of the group LASSO, the randomized group LASSO. Applied to a subsample k, the randomized group LASSO estimates $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}^{\text{grp}}(\lambda; \mathcal{D}^{(k)}) = \underset{\boldsymbol{\beta}_a, \boldsymbol{\beta}_d}{\text{argmin}} \left\{ -\ell(\boldsymbol{\beta}_a, \boldsymbol{\beta}_d; \mathcal{D}^{(k)}) + \lambda \sum_{j=1}^{m} \frac{1}{U_j} \sqrt{{\beta_{a,j}}^2 + {\beta_{d,j}}^2} \right\}, \tag{3.5}$$

where $U_j \sim (\alpha, 1)$ is a weighting parameter with $\alpha \in [0, 1]$. U_j randomly down-weights some predictors relative to others. We note that the the randomized group LASSO can be easily implemented with any group LASSO software by scaling the predictors within each group by the generated weight U_j. The performance of the randomized group LASSO is dependent on the value of α used. We calibrate the randomization parameter based on simulations, see Results.

When using a randomized selection procedure, one must consider more closely the number of resamples K to run. With the additional variability in the estimation of $\boldsymbol{\beta}$, which is dependent on both α and the distribution used for generating U_i's, it is important to perform sufficient resamples to compensate for the additional variability

in the RMIP. Meinshausen and Bühlmann (2010) proposed to generate U_i's as either α or 1 with equal probability for the randomized LASSO. We have found that this method of generating U_i's works well with the randomized group LASSO, and hat the choice of α is robust with respect to the model choice. Our results suggest that 250 resamples is sufficient when using $\alpha = 0.7$. This is a 2.5 fold increase to the number of resamples sufficient for RMIPs to converge when using the standard LASSO, i.e. no randomization (Valdar et al., 2012).

3.2.3 Competing methods

Our RMA generalization of LLARRMA calculates a score (an RMIP) for each SNP in the identified hit region. We compare the ability of those scores to discriminate true signals from background SNPs with the SNP scores calculated by two alternatives: the traditional GWAS approach of single locus regression, and LLARRMA (Valdar et al., 2012).

Single locus regression

We perform single locus regression with linear regression as used in, for example, PLINK (Purcell et al., 2007). For each SNP, we fit a single locus dominance model version of Eq 3.1 and score its $-\log_{10} P$ ("logP"), where P is the p-value from a likelihood ratio test against an intercept-only model. We also compare with with the additive only single locus regression model.

LLARRMA

We compare the generalized RMA procedure with the original additive only LLARRMA procedure. We also use the standard LLARRMA procedure with only some of the generalizations proposed as intermediate steps to our full proposed method, see Terminology used in the paper for considered variations.

3.2.4 Terminology used

We describe our decomposition of the proposed generalization of LLARRMA (Valdar et al., 2012) based on a character abbreviation describing the assumed model and resampling method. Table 3.1 describes this coding. The coding describes the choices one may make among our generalizations: the type of resampling, the typed of modelable effects, and if locus predictors are to be group for non-additive models. For example, the simplest RMA procedure used is LLARRMA, a subsampling-based RMA with an additive only model. We abbreviate this model as LLARRMA-as where the 'a' is for additive model and the 's' is for subsampling. Generalizing the additive model in LLARRMA to account for dominance predictors, still using the LASSO, would be notated as LLARRMA-das, where the 'd' is for the addition of dominance into the model. To incorporate the sample weighting, we would replace the 's' with a 'w', where 'w' is for weighted resamples. When considering sample weighting, we may add a 'g' to indicate the use of the group LASSO on the general model which allows for dominance effects. For example, our full proposed method, weighted resampling RMA with the group LASSO, would be called LLARRMA-dawg.

Table 3.1: Nomenclature for modeling and resampling procedures used in the paper.

Character	Description
s	Resampling by subsampling
w	Resampling by weighted samples
a	Additive effects
d	Dominance/Heterosis effects
g	'a' and 'd' modeled as a predictor group

3.3 Simulation framework

In order to evaluate the extensions of the LLARRMA framework to incorporate dominance modeling, we consider a variety of settings to test the models. We present two simulations studies. The first will focus on method performances when considering a

single effect type, i.e. only additive, only minor allele dominant, etc.. The second considers models which contain a mixture of effect types. The first will focus on a model which effects are most likely to be additive, while the second will focus on the case where additive effects are less prevalent.

3.3.1 Simulating Genotypes

We have chosen to simulate data sets using HAPGEN2 (Su, Marchini and Donnelly, 2011) which was developed for generation of SNP data for complex diseases that have an LD structure mimicking a provided real data set. With the use of HAPGEN2, we are able to easily generate a new data set for each simulation, allowing us to test on a wider variety of data sets than if we generated phenotypes based on a single fixed real data set.

The simulated data set we consider is a HAPGEN2 version of the hit region from the '58 data (WTCCC, 2007) used in Valdar et al. (2012). As not all of the SNPs selected from the '58data are available from HapMap (Tanaka, 2009); the subset of 386 SNPs of the hit region present in HapMap are used to generate our HAPGEN2 data set. Each data set will consist of 2500 subjects. The LD of this hit region is displayed in Fig 3.1.

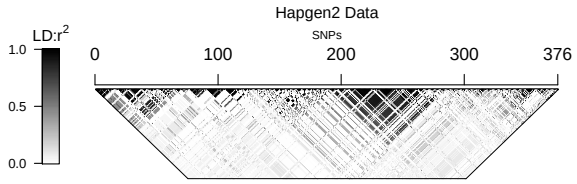

Figure 3.1: LD structure of the HAPGEN2 data sets used in the simulations. Shading indicates pairwise LD between SNPs, ranging from white ($r^2 = 0$) to black ($r^2 = 1$).

3.3.2 Simulation study 1: preliminary model comparisons

When extending the RMA model to detect dominance in the model we want to compare how the extended model will fair in multiple different settings. Each sub-simulation is

performed to evaluate how each method compares when used on a model with only a single type of effect.

Placement of true loci

The location of the true loci for the simulation sub-studies will be chosen at random, with a restriction on the minor allele frequencies (MAF) of the selected loci. The MAF of true signal SNPs has been restricted to be at least 0.1 to ensure sufficient signal is present to detect dominant effects.

Simulating phenotypes

Phenotypes are simulated based on the regression model given by Eq 3.1. Given a set of true SNPs with genotypes $\mathbf{X}_q$ with corresponding model predictors $\mathbf{A}_q$ and $\mathbf{D}_q$ and their corresponding effects $\boldsymbol{\beta}_{a,q}$ and $\boldsymbol{\beta}_{d,q}$, we first calculate individuals expected phenotype $y_i = \mathbf{A}_q{}^{\mathrm{T}}\boldsymbol{\beta}_{a,q} + \mathbf{D}_q{}^{\mathrm{T}}\boldsymbol{\beta}_{d,q}$, and then add a Gaussian error $e_i \sim N(0, \sigma)$ to obtain the individual's observed phenotype, where σ is chosen to obtain the desired signal to noise ratio (SNR) of 1/4, where SNR $= \frac{\sqrt{[\boldsymbol{\beta}_a, \boldsymbol{\beta}_d]^T \mathrm{var}([A,D])[\boldsymbol{\beta}_a, \boldsymbol{\beta}_d]}}{\sigma}$. This corresponds to the region explaining 5.8% of the phenotypes variability, which is comparable to the observed variability explained within hit regions in Warren et al. (2012) and Dastani et al. (2012).

Simulation substudies: generation of model effects

The simulations are broken into 5 sub-simulations in order to investigate how each model performs in each of the specific settings. For each sub-simulation we will consider 5 true loci with effects $\boldsymbol{\beta}_q^\star$ generated as $\mathrm{N}(1.35(-1)^{\nu_j}, 0.02^2)$ with $\nu_j \sim$ Bernoulli(0.5). Each sub-simulation will emphasize a different combination of $\boldsymbol{\beta_a}$ and $\boldsymbol{\beta_d}$ as a function of $\boldsymbol{\beta}_q^\star$ to consider additive only, heterosis, and general dominant effects. Table 3.2 summarizes the settings for each sub-simulation.

Table 3.2: Summary of the sub-simulation models where $\boldsymbol{\beta}_q^\star \sim \text{N}(1.35(-1)^{\nu_j}, 0.02^2)$ with $\nu_j \sim$ Bernoulli(0.5), α is chosen randomly from $\{0.5, 0.75, 1, 1.25\}$, and $\upsilon_j \sim$ Bernoulli(0.5).

Substudy	Model	Additive predictor	Dominant predictor
1A	Additive	$\boldsymbol{\beta}_{a,q} = \boldsymbol{\beta}_q^\star$	$\boldsymbol{\beta}_{d,q} = \mathbf{0}$
1B	Minor Allele Dominant	$\boldsymbol{\beta}_{a,q} = \boldsymbol{\beta}_q^\star$	$\boldsymbol{\beta}_{d,q} = \boldsymbol{\beta}_{a,q}$
1C	Major Allele Dominant	$\boldsymbol{\beta}_{a,q} = \boldsymbol{\beta}_q^\star$	$\boldsymbol{\beta}_{d,q} = -\boldsymbol{\beta}_{a,q}$
1D	Heterosis	$\boldsymbol{\beta}_{a,q} = \mathbf{0}$	$\boldsymbol{\beta}_{d,q} = \boldsymbol{\beta}_q^\star$
1E	General Dominant	$\boldsymbol{\beta}_{a,q} = \boldsymbol{\beta}_q^\star$	$\boldsymbol{\beta}_{d,q} = \alpha(-1)^{\upsilon_j}\boldsymbol{\beta}_{a,q}$

3.3.3 Simulation study 2: general predictors

For our second simulation study, we will consider a general setting that will be a mixture of the effects tested in simulation study 1. Specifically we will consider a combination of simulation 1A, 1D, and 1E; as the settings of 1B and 1C are special cases of 1D.

Let (m_a, m_d, m_h) be the number of true additive, dominant, and heterosis effects in the model respectively. We propose to model $(m_a, m_d, m_h) \sim \text{Multinomial}(5, p_a, p_d, p_h)$ where $p_a, p_d,$ and p_h are the probabilities of a true locus being additive, dominant, and heterosis effects respectively. Under this model, we can characterize the simulation 1 sub-studies by explicitly setting two of the three probabilities to zero. Thus, we have a natural extension of the simple sub-studies to a more general simulation. We propose two simulation settings. The first models $(m_a, m_d, m_h) \sim \text{Multinomial}(5, p_a = .6, p_d = .3, p_h = .1)$, which deviates slightly from the standard complex trait analysis assumption by allowing some effects to differ from additivity. In the second setting, we model $(m_a, m_d, m_h) \sim \text{Multinomial}(5, p_a = .3, p_d = .6, p_h = .1)$, which emphasizes a more extreme view where the non additive effects are most prevalent.

3.3.4 Computation

All analyses were performed in R (R Development Core Team, 2010), with the *glmnet* package (Friedman, Hastie and Tibshirani, 2010) used for fitting LASSO models and the *grplasso* package (Meier, 2009) for group LASSO models.

3.4 Results

3.4.1 Calibrating the randomization penalty

We explore the use of RMA under two types of penalties: standard penalty, i.e. under constant penalization, and 'randomized' penalty, i.e. under random perturbation of the predictors' penalization. The purpose of the randomized penalty is to perturb the level of penalization on individual predictors to address the 'instability' of the LASSO or group LASSO with highly correlated predictors. The degree of perturbation is controlled by the randomization parameter α. In the paper that introduces the randomized LASSO, Meinshausen and Bühlmann (2010) advocated choosing $\alpha \in [0.2, 0.8]$, stating that there was little change in the performance of the procedure within this region. Although our findings based on the full AUC are consistent with their findings, we found that the performance based on initial AUC differs based on the value of α. We also found that the choice of randomization parameter α is dependent on the data's correlation structure; how the random perturbation of predictor penalties effects a method may depend on the relationship between the variables (i.e. correlation or LD). We have found that under simulations of SNP hit regions based on our data choosing $\alpha \in [0.6, 0.8]$ gave optimal performance based on initial AUC, and so we set $\alpha = 0.7$ throughout.

3.4.2 An Example Simulation

Figure 3.2 plots SNP location against SNP-score for select methods in an example simulation. True signal SNPs are plotted as black crosses and the remaining (background) SNPs as gray dots. In single-locus regression (Figure 3.2A), SNPs are scored as -log10 P (logP; see Methods). Although the true signals between 150 and 200 tend to attract higher scores, so do many of the backgrounds SNPs, giving rise to a cloud of association that is characteristic of many hit regions in real GWAS. The remaining methods (3.2

B–D) report inclusion probabilities (RMIPs) for each SNP. Figure 3.2B displays the standard LLARRMA (Chapter 2) output. Figure 3.2C displays the output of LLARRMA when including the weighted resamples and group LASSO generalizations we propose. Figure 3.2D incorporates the calibrated randomized group LASSO. In this example, all RMA based methods are enriched for true signals when compared with single-locus regression. We observe that our proposed generalizations further enrich the output from standard LLARRMA.

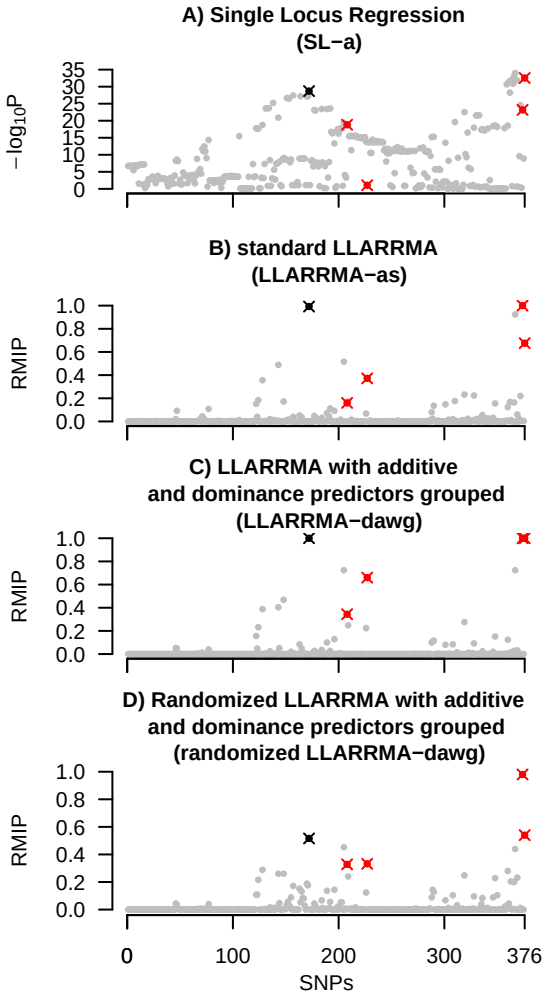

Figure 3.2: Results of four methods applied to an example dataset from simulation study 2B. Plots show SNP score (logP or RMIP) against SNP location in the Hapgen2 data, with true signal SNPs in black (additive effect) and red (non-additive effect) and background SNPs in gray.

3.4.3 Simulation study 1: individual effect types

To examine the relative performance of the single and multiple locus methods, we simulated 500 data sets based on the Hapgen2 data (see Methods and Figure 2.1). Each simulated dataset had a complete set of genotypes for 2500 individuals. Their outcomes were influenced by 5 true SNPs of moderate effects (SNR=1/4 or 5.8% of variability explained), with 500 simulations devoted to each type of true signals from Methods. For each simulation we tested three different analysis methods with varying model types (additive or dominant) that each produced a score per SNP. Our subsequent comparisons of those methods were based on how well their scores discriminated the five true signal SNPs from the background SNPs. All LLARRMA-based methods (i.e., all except single locus regression) used K = 250 resamples and both their standard versions (i.e. LASSO and group LASSO) and their randomized versions (i.e. randomized LASSO and randomized group LASSO).

Results from 500 simulations

Figure 3.3 plots ROC curves (see Methods) for each method in each sub-setting. The ROC curve plots the trade-off between power (the proportion of true signals declared as influential) and FPR (the proportion of background SNPs declared as influential) when thresholding the SNP scores (logPs or RMIPs). The initial ROC is arguably of greater relevance to GWAS because it focuses on enrichment of true signals among the top-scoring SNPs. A method whose top four SNPs are true signals, but which never finds the fifth true signal SNP, is arguably more valuable than one whose top SNPs are false but which finds all five true signals among its middle scoring SNPs (Chapter 2). Figure 3.3 shows both the full ROC curve and the initial ROC curve for each substudy. Figure 3.3 focuses on the difference between LLARRMA-s and LLARRMA-w procedures. Although there are some apparent advantage to the LLARRMA-w variations based on visual comparisons, examining the AUCs emphasizes the improvements. The

use of the randomized version of the LASSO and group LASSO consistently improves performance, as can be seen in Table 3.3. Table 3.3 displays the mean percent of initial AUCs from simulation study 1 with the best AUC for each substudy bolded. Within both the standard and randomized procedures, we observe a slight advantage to weighted resampling over subsampling, and see that there is an advantage to the randomized procedures for each RMA variation. This suggests that both the use of weighted resampling and the randomized group LASSO have an advantage over standard RMA procedure. The results also suggest strong advantages to simultaneously modeling multiple loci in the presence of high LD.

Table 3.3: Mean percent of maximum initial AUC for simulation study 1. All standard errors are less than 0.94. Bold indicates the best method for each model and any methods statistically the same as the best method. Underlined indicates the best method excluding randomized procedures and any methods statistically the same as the best non-radomized method.

Simulated Model	Single locus regression		LLARRMA Standard Penalization				
	-a	-da	-as	-das	-aw	-daw	-dawg
Additive	16.1	16.1	61.7	<u>63.1</u>	61.4	<u>62.7</u>	60.9
Minor Dom.	14.9	16.8	48.2	52.6	48.7	52.9	<u>54.4</u>
Major Dom.	19.9	20.5	46.5	51.6	46.8	53.4	<u>58.0</u>
Heterosis	14.9	17.8	49.4	54.8	49.0	54.6	<u>56.7</u>
General Dom.	11.8	19.3	41.2	<u>77.4</u>	41.6	75.8	75.6
Simulated Model	Single locus regression		LLARRMA Randomized Penalization				
	-a	-da	-as	-das	-aw	-daw	-dawg
Additive	16.8	16.8	63.2	**64.1**	63.6	**64.4**	63.2
Minor Dom.	15.3	16.9	50.5	54.1	51.1	54.9	**57.5**
Major Dom.	19.8	20.4	48.0	51.9	48.9	54.5	**59.8**
Heterosis	14.6	17.6	51.4	55.9	51.3	56.0	**58.9**
General Dom.	11.3	17.9	43.3	77.6	43.8	77.6	**78.3**

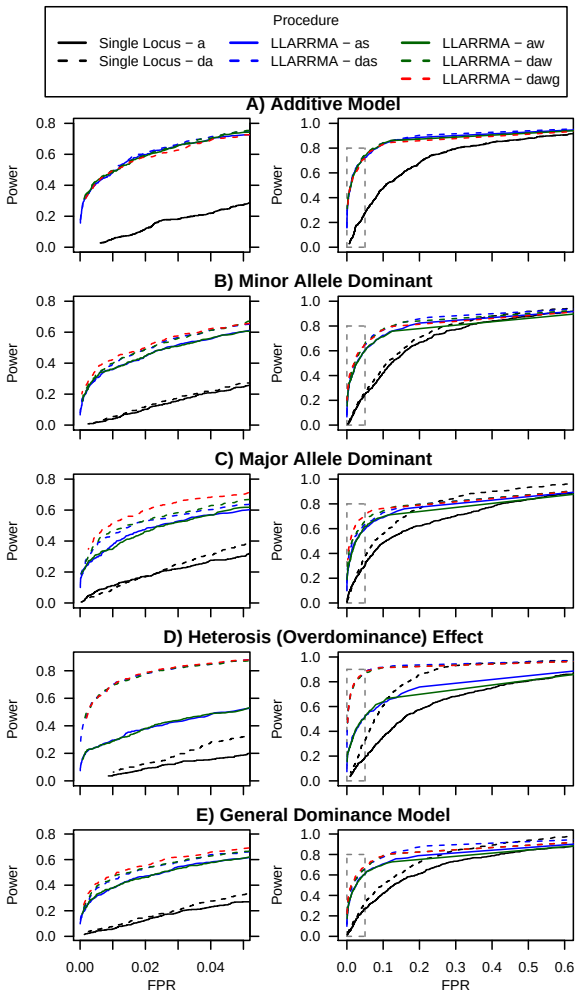

Figure 3.3: Initial and full ROC curves for simulations study 1's 5 sub-studies. We observe an overwhelming difference between the single locus and multiple locus methods in all situations. We observe consistently that LLARRMA-w procedures perform at least as well as there LLARRMA-s counterparts.

Comparing SL and RMA SNP ranking for General Dominance

Figure 3.4 plots SL rank against RMA based method's rank of the 5 true signals in each of the 500 simulations in study 1E. True signals are plotted based on their significance based on SL logP's: genome wide significant (logP $\geq$ 8; green), marginal significance (orange), not significant (logP $\leq$ 4; red). We observe that RMA methods are able to identify a significant number of true signals that failed to reach even marginal significance based on SL scans. We observe that our generalizations of

LLARRMA (LLARRMA-dawg) has improved performance from standard LLARRMA (LLARRMA-a). We also observe a large improvement with the addition of the randomized group LASSO (randomized LLARRMA-dawg).

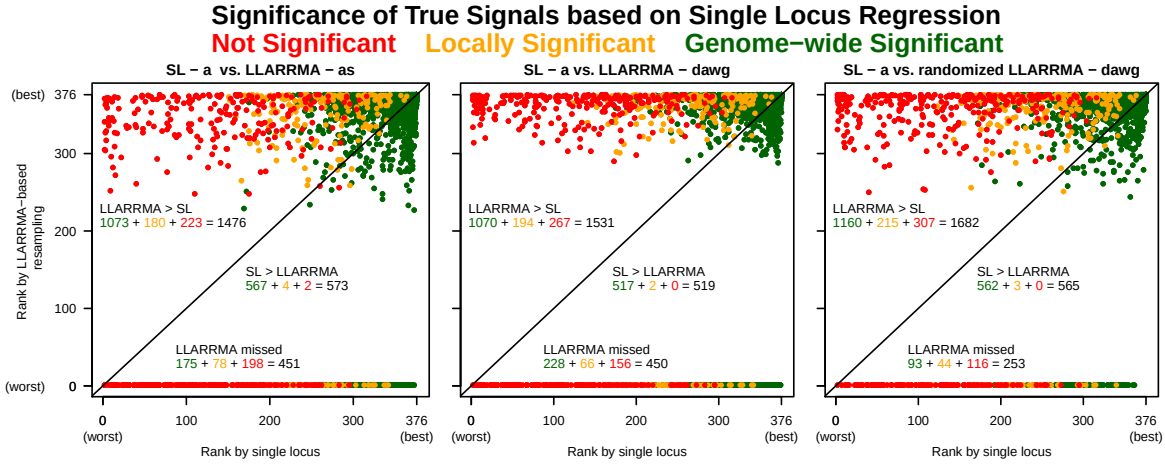

Figure 3.4: Ranking of 2500 true signals from study 1E by single locus regression (SL) vs by LLARRMA-based method (RMA). Colors based on SL significance; genome wide significant (logP ≥ 8; green), marginal significance (orange), not significant (logP ≤ 4; red).

To further compare how each method compares at finding true signals, we calculate the average number of SNPs that must be examined to find the true signals for each method. Figure 3.5 plots the average number of SNPs that must be examined to find the first, second, third, fourth, and fifth true signal in simulation 1E. We observe very little difference when considering only the first true signal, but single locus regression quickly takes significantly more SNPs to find the second SNP than RMA based methods. We observe a consistent improvement from methods which allow for dominance effects. The performance between corresponding subsampling based and weighted resampling based methods appear very similar, indicating that the technical motivated generalization of weighted resampling has little effect on performance.

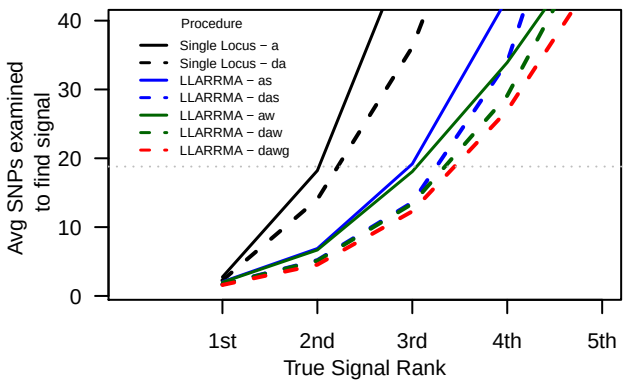

Figure 3.5: The average number of SNPs that must be examined to find the first, second, third, fourth, and fifth true signal in simulation 1E. Dotted gray line indicates 5% of the SNPs in the hit region.

3.4.4 Simulation study 2: general effects

We simulated 500 datasets from HAPGEN2 for each setting. Each dataset is influenced by 5 true loci with moderate SNP effects (SNR=1/4 or 5.8% of variability explained). The number of additive, general dominant, and heterosis predictors (m_a, m_d and m_h respectively) followed a Multinomial$(5, p_a, p_d, p_h)$ distribution where for study 2A we select parameter values that focus on mostly additive effects. In study 2B, we adjust the parameters so that the effects are more focused on non-additive effects.

Results from 500 simulations from study 2A

Figure 3.6A plots the initial and full ROC curves from the 100 simulations in which the 5 causal loci were characterized by a Multinomial$(5, p_a = 0.6, p_d = 0.3, p_h = 0.1)$ distribution. We observe all forms of randomized LLARRMA dominate single locus regression in the initial ROC curve, suggesting an advantage of simultaneously modeling multiple loci in the presence of high LD. We also notice that the dominant models are superior to additive only models, which is consistent with the findings from simulation study 1. The findings are also consistent with study 1 in that the LLARRMA-dawg outperformed all methods in the initial ROC.

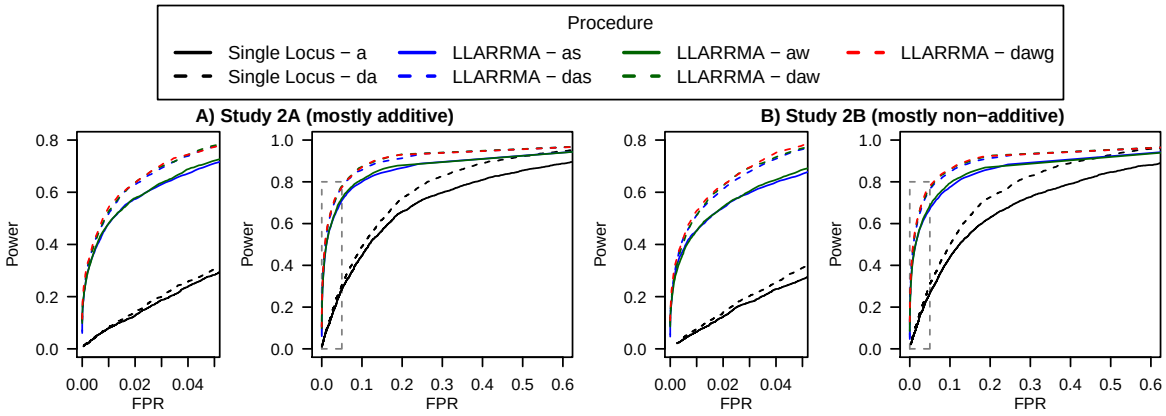

Figure 3.6: Initial and full ROC curves for simulations study 2A ($p_a = 0.6, p_d = 0.3$, and $p_h = 0.1$) and 2B ($p_a = 0.3, p_d = 0.6, p_h = 0.1$). All LLARRMA procedures are using their randomized penalties.

Results from 500 simulations from study 2B

Figure 3.6B plots the initial and full ROC curves from the 500 simulations in which the 5 causal loci were characterized by a Multinomial$(5, p_a = 0.3, p_d = 0.6, p_h = 0.1)$ distribution. In this setting, we see a similar trend to that of simulation study 1 and simulation study 2A.

Table 3.4 displays the mean initial AUC from simulations studies 2A and 2B. We observe a clear advantage of the randomized LLARRMA-dawg procedure proposed here over all other methods in each setting. The results are consistent with those of simulation study 1 in that the further from additive the true model is, the greater the advantage LLARRMA-dawg (or randomized LLARRMA-dawg) has.

3.5 Theory: Bounds on false positives

This section discusses the modification of a bound on the expected number of false positives for stability selection (Meinshausen and Bühlmann, 2010). Before discussing our theorems, consider more formal definitions of previously used terminology. Consider vector valued data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ which can be taken to be a realization of IID random elements $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in \mathcal{R}^p$. Assume that some of the components of $\mathbf{Z}_i$ are 'signal

Table 3.4: Mean percent of total initial AUC for simulation study 2, where in 2A the true signals effect types are sampled from a Multinomial$(5, p_a = 0.6, p_d = 0.3, p_h = 0.1)$ and 2B from a Multinomial$(5, p_a = 0.3, p_d = 0.6, p_h = 0.1)$ distribution. All standard errors are less than 0.92. Bold indicates the best method for each model and any methods statistically the same as the best method. Underlined indicates the best method excluding randomized procedures and any methods statistically the same as the best non-radomized method.

Simulated	Single locus regression		LLARRMA Standard Penalization				
Model	-a	-da	-as	-das	-aw	-daw	-dawg
Mostly add. (2A)	15.6	16.8	54.9	<u>61.0</u>	54.9	<u>60.6</u>	<u>60.7</u>
Mostly non-add. (2B)	14.6	16.5	51.9	<u>59.7</u>	52.0	<u>59.9</u>	<u>59.6</u>

Simulated	Single locus regression		LLARRMA Randomized Penalization				
Model	-a	-da	-as	-das	-aw	-daw	-dawg
Mostly add. (2A)	15.6	16.8	56.9	62.4	57.5	62.7	**63.3**
Mostly non-add. (2B)	14.6	16.5	54.1	61.3	54.5	61.9	**62.9**

variables', and the others are 'noise variables'. Formally, define the set $S \subseteq \{1, \ldots, p\}$ and the set $N = \{1, \ldots, p\} \backslash S$ to be the 'signal' and 'noise' variables respectively.

Definition 1 (Variable selection procedure)**.** *A variable selection procedure is a statistic* $\hat{S}_n = \hat{S}_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ *taking values in the set of all subsets of* $\{1, \ldots, p\}$*. Consider* $\hat{S}_n$ *as an estimate of* S*. Further, define the expected number of variables included by* $\hat{S}_n$ *as* $q_n = E(|\hat{S}_n|)$*.*

Definition 2 (Weighted variable selection procedure)**.** *A weighted variable selection procedure is a variable selection procedure that is capable of accepting a set of observation weights,* $\mathbf{W} = \{w_1, \ldots, w_n\}$*. Formally, it is a statistic* $\hat{S}_{n,w} = \hat{S}_{n,w}(Z_1, \ldots, Z_n, \mathbf{W})$ *taking values in the set of all subsets of* $\{1, \ldots, p\}$*. Consider* $\hat{S}_{n,w}$ *as an estimate of* S*. Further, define the expected number of variables included by* $\hat{S}_{n,w}$ *as* $q_n = E(|\hat{S}_{n,w}|)$*.*

Definition 3 (Model Inclusion Probability (MIP))**.** *We define the MIP of a variable index* $k \in \{1, \ldots, p\}$ *under* $\hat{S}_n$ *(or* $\hat{S}_{n,w}$*) as*

$$\Pi_{k,n} = P(k \in \hat{S}_n).$$

Note that we will assume that n is fixed based on the sample, and that w is clearly defined based on the assumed weighting function and will drop the n and w subscripts from here on out except for where they are clearly needed.

Definition 4 (Generalized Resample Model Averaging (RMA))**.** *Let* $\{\mathbf{W}_b : b = 1, \ldots, B\}$ *be a set of randomly chosen independent weights such that* $\mathbf{W}_b \in [0,1]^n$ *is from a weighting function with* $E(\mathbf{W}_b) = \frac{1}{2}\mathbf{1}$ *for all* b*. Consider a weighted variable selection procedure* $\hat{S}$*. For* $\tau \in [0,1]$*, the generalized RMA version of* $\hat{S}$ *is* $\hat{S}_\tau^{\text{RMA}} = \{k : \hat{\Pi}_k \geq \tau\}$*, where*

$$\hat{\Pi}_k = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W}_b)\}}$$

Definition 5 (Complement weight)**.** *The complement to a weight* $\mathbf{W} \in [0,1]^n$ *is defined to be* $\mathbf{1} - \mathbf{W}$*.*

Definition 6 (Simultaneous selection probability)**.** *Let* $\{\mathbf{W}_b : b = 1, \ldots, B\}$ *be randomly chosen independent weights such that* $\mathbf{W}_b \in [0,1]^n$ *with* $E(\mathbf{W}_b) = \frac{1}{2}\mathbf{1}$ *for all* b*. Let* $\{\mathbf{W}_b^\star : b = 1, \ldots, B\}$ *be the set of complement weights to* $\mathbf{W}_b$*, i.e.,* $\mathbf{W}_b^\star = \mathbf{1} - \mathbf{W}_b$*. For* $\tau \in [0,1]$*, the simultaneous selection version of* $\hat{S}$ *is* $\hat{S}_\tau^{\text{simult}} = \{k : \hat{\Pi}_k^{\text{simult}} \geq \tau\}$*, where*

$$\hat{\Pi}_k^{\text{simult}} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W}_b)\}} \mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W}_b^\star)\}}$$

Results for RMA

With the notation defined above, we can now state the theorem for the bound of the expected number of falsely selected variables under an RMA procedure (i.e. generalized RMA with subsample weighting).

Theorem 3.1 (Error Control for Resample Model Averaging)**.** *Assume that the weighting function in the generalized Resample Model Averaging algorithm is the subsampling(.5) function and that the distribution of* $\{\mathbb{I}_{\{k\in\hat{S}\}}, k \in N\}$ *is exchangeable. Suppose that the selection procedure is not worse than random guessing, i.e.,*

$$\frac{E(|S \cap \hat{S}|)}{E(|N \cap \hat{S}|)} \geq \frac{|S|}{|N|}.$$

Let $q = E(|\hat{S}|)$ *be the average number of variables selected on a given resample. Then, the expected number of falsely selected variables,* V*, with* RMIP $\geq \pi_{\text{thr}} \in (\frac{1}{2}, 1)$ *is bounded by*

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1}\frac{q^2}{p}.$$

We have thus established the analogous bound from stability selection for the more general RMA framework. This bound, like with that for SS's bound, is very general and will hold for any variable selection procedure that is better than random guessing. In many cases, e.g. permutation selection in LLARRMA, the variable selection method is substantially better than random guessing. This results in a bound that is too liberal. When we are able to establish how much better than random guessing a variable selection method is, we are able to obtain the improved bound stated below.

Theorem 3.2 (Improved Error Control for Resample Model Averaging)**.** *Assume that the weighting function in the generalized Resample Model Averaging algorithm is the subsampling(.5) function and that the distribution of* $\{\mathbb{I}_{\{k\in\hat{S}\}}, k \in N\}$ *is exchangeable.*

Suppose that the selection procedure is not worse than γ times random guessing, i.e.,

$$\frac{E(|S \cap \hat{S}|)}{E(|N \cap \hat{S}|)} \geq \gamma \frac{|S|}{|N|}.$$

Let $q = E(|\hat{S}|)$ be the average number of variables selected on a given resample. Then, the expected number of falsely selected variables, V, with RMIP $\geq \pi_{\text{thr}} \in (\frac{1}{2}, 1)$ *is bounded by*

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q^2 p}{(p + (\gamma - 1)|S|)^2}.$$

The amount that the result above is able to improve on the original bound is dependent on two things. These are γ (how much better than random guessing the variable selection method is) and the number of true variables. We can assume that $|S| \geq 1$ in our setting as we assume that the regions we analyze have been preselected as regions with multiple loci associated with the response. The better we are able to estimate the number of true signals present, the more we are able to improve the bound. In practice, it may be easy to get a low bound on the number of true signals, but a good estimation of γ may be difficult.

Results for generalized RMA

Under the more general setting described in this chapter, it is not possible to obtain the same theoretical bound for generalized RMA which uses the LASSO under the same assumptions. Considering a general weighting function loses some of the nice properties that we had under subsampling weights. Specifically, the proofs of Theorem 3.1 and 3.2 (See Appendix A) utilize the conditional independence of the subsample and complement sample conditioned on the observed data. Under a general weighting function, without further assumptions, and obtains the following bound given by Theorem 3.3.

Theorem 3.3 (Error Control for Generalized Resample Model Averaging)**.** *Assume a symmetric weighting function in the generalized Resample Model Averaging algorithm*

and that the distribution of $\{\mathbb{I}_{\{k\in\hat{S}\}}, k \in N\}$ *is exchangeable. Suppose that the selection procedure is not worse than* γ *times random guessing, i.e.,*

$$\frac{E(|S \cap \hat{S}|)}{E(|N \cap \hat{S}|)} \geq \gamma \frac{|S|}{|N|}.$$

Let $q = E(|\hat{S}|)$ *be the average number of variables selected on a given resample. The expected number of falsely selected variables,* V*, with* RMIP $\geq \pi_{\text{thr}} \in (\frac{1}{2}, 1)$ *is bounded by*

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{pq}{(p + (\gamma - 1)|S|)}.$$

The bound established in Theorem 3.3 is essentially the most liberal bound. We would like to be able to further improve the bound. By making the further assumption that the variable selection procedure used has a monotonicity property of inclusion of null variables with the respect to the weights varying (i.e., the function $\mathbb{I}_{\{k\in\hat{S}(\mathbf{Z},\mathbf{W})\}}$ is monotone with respect to $\mathbf{W}$ for a null variable k), one can appeal to a covariance type inequality to obtain the bound we obtained under subsampling. In general, the LASSO does not satisfy this property. An example of a procedure that would meet this requirement is the LASSO applied to independent data. The bound for generalized RMA with a monotone selection procedure in Theorem 3.4.

Theorem 3.4 (Error Control for Generalized Resample Model Averaging with a monotone selection procedure)**.** *Assume a symmetric weighting function in the generalized Resample Model Averaging algorithm using a monotone selection procedure and that the distribution of* $\{\mathbb{I}_{\{k\in\hat{S}\}}, k \in N\}$ *is exchangeable. Suppose that the selection procedure is not worse than* γ *times random guessing, i.e.,*

$$\frac{E(|S \cap \hat{S}|)}{E(|N \cap \hat{S}|)} \geq \gamma \frac{|S|}{|N|}.$$

Let $q = E(|\hat{S}|)$ be the average number of variables selected on a given resample. Then the expected number of falsely selected variables, V, with RMIP $\geq \pi_{\text{thr}} \in (\frac{1}{2}, 1)$ *is bounded by $\pi_{\text{thr}} \in (\frac{1}{2}, 1)$ by*

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{pq^2}{(p + (\gamma - 1)|S|)^2}.$$

3.6 Discussion

We present general extensions of previous RMA works for characterizing frequentist variability in model selection and apply them to a well suited problem, discriminating true signals from false signals among a set of SNPs that are often highly correlated. Single locus methods often fail to discriminate true signals from background noise in hit regions of high LD. We re-emphasize that that multiple locus based procedures, e.g. RMA, outperform single locus regression in this setting (Valdar et al., 2012). While doing so, we also describe inadequacies of additive only genetic models. Specifically, we extend the genetic model to include the ability to modeling more general genetic effects such as dominance. We propose the group LASSO for variable selection on a locus level when considering the more general model, and lose little performance when the true model is in fact additive. In doing so, we exposed a technical issue of subsampling based RMA; when considering rarer predictors such as the dominance predictor here, or more generally rare variants themselves, a predictor may become monomorphic as a result of the subsampling. To address this issue, we replace subsampling with weighted resampling.

The use of weighted resampling results in a more robust procedure. Rare predictors, e.g. dominance predictors or rare variants, can be problematic for subsampling as the subsampling can unintentionally remove predictors if they become monomorphic (i.e., a predictor or locus becomes constant, providing no useful information to

the model) on a given subsample. The generalization to weighted resamples gives each subject some weight in the fitting of each resample, ensuring that no predictor becomes monomorphic. We observe that in a setting where it is unlikely for subsampling to observe monomorphic predictors (e.g. when considering an additive model without rare variants) that the weighted resampling has nearly identical performance. When considering the more general model (e.g. allowing dominance predictors) we observe a slight increase in overall performance when using weighted resampling. This generalization is similar in spirit to how the Bayesian bootstrap generalizes the standard bootstrap (Rubin, 1981). That is, our weighted resampling generalizes the corresponding discrete 0 or 1 weights of subsampling to continuous weights in the same way the Bayesian bootstrap generalizes the discrete multinomial weights of a bootstrap to a continuos dirichlet weight. The Bayesian bootstrap is a specific example of the weighted likelihood bootstrap (Newton and Raftery, 1994), which was shown to provide a sampling distribution of the statistic of interest that at least has the correct first moment. Our proposed weighting distribution may also be represented under the weighted likelihood bootstrap formulation. We observe that when considering the simple example of estimating the sample mean, our proposed weighting results in a sampling distribution of the sample mean that is consistent with the use of subsampling or the bootstrap.

When considering weights $w_{i,k} \in [0, 1]$, the choice of weighting($\cdot$)=subsampling(ϕ) provides one extreme of the values that the weights can take (i.e., weights $w_{i,k} \in \{0, 1\}$). The other extreme would be to consider a function for weighting($\cdot$)that returns the same weight for each observation, which would be consistent with not resampling and just fitting the model on the entire data set once. Our choice of weighting($\cdot$) =U(0,1) provides a natural midpoint between the two extremes one may use for their weights.

The assumption of only additive SNP effect has been an effective statistical simplification of a more complex genetic model. If the underling genetic model is not additive,

the ability to model dominant effects, or some deviation from the additive effect at the heterozygote, can increase power to detect true signals. We generalize the additive only model of our previous work, LLARRMA (Valdar et al., 2012), to include additional predictors which allow us to model more general SNP effects such as dominance. We observed that the generalization of the model does not hinder performance when the underling model is additive. When the underlying model differs from additivity, we observe a significant increase in performance when considering the more general model. When extending the additive genetic model of our previous work to model more complex genetic effects, we found the group LASSO to be an intuitive modification for variable selection. By grouping all model predictors for a single locus together, the group LASSO allows us to address our main goal, to identify which loci have true associations with the phenotype. Unfortunately, the group LASSO does not allow for the distinction between additive and dominant effects. It may be possible to modify the group LASSO penalty in order to distinguish between additive and dominant loci effects.

Our findings suggest that when using a RMA based procedures, the use of the randomized penalties (i.e the randomized group LASSO or LASSO) improves the RMA procedure's ability to discriminate true signals from background SNPs. We show that randomized penalties can to be quite powerful for correlated data when incorporated with resampling. The randomized group LASSO (or LASSO) is highly dependent on both the value of the randomization parameter α, and may require calibration of α. In our calibration, we found that the value of α had little effect on the performance when we considered the full AUC. This was consistent with the findings of Meinshausen and Bühlmann (2010) who advocated choosing $\alpha \in [0.2, 0.8]$ for the randomized LASSO, stating that there was little change in the performance of the procedure within this region. What found that the value of α does have a larger impact on the initial AUC.

Specifically, as we decreased α past a point, the average initial AUC decreases past the AUC which you would have obtained if you had not used the randomized version. Our calibrations found that $\alpha \in [0.6, 0.8]$ had similar performance with the maximum average initial AUC when $\alpha = 0.7$. Further examination is still needed to see how far these values may generalize. We have also found that the randomized procedure requires a larger number of resamples for convergence of the RMIP estimates.

Resample aggregation techniques such as bootstrap aggregation ("bagging"; Breiman, 1996) or subsample aggregation (subagging; Bühlmann and Yu, 2002) have been found to produce estimates of $\boldsymbol{\gamma}$ that are more stable than estimates obtained from a single model selection. Specifically, the aggregation estimates have lower frequentist risk under squared error loss (Bühlmann and Yu, 2002). The addition of a RMA-w option, a bagging based approach, within the resample model averaging framework (Valdar et al., 2012) gives option of using either a bagging or subagging based resampling approach when a particular problem may be more suited for one method, such as using RMA-w when considering dominance or rare alleles.

Although we describe LLARRMA-dawg in the quantitative setting of a linear model, it is easily extended to logistic regression for case/control phenotypes, or any generalized linear model for which the LASSO or group LASSO may be applied. Similarly, while we described how the group LASSO may be incorporated to expand the scope of additive SNP effects to dominant effects, a similar modification may be done to examine local haplotype effects.

In summary, we describe multiple modifications to RMA for characterizing frequentist variability of model choice that can be usefully applied to SNPs in hit regions of a GWAS. The authors will provide an implementation of the proposed modifications to LLARRMA in the R-package R/llarrma as soon as is practicable.

Chapter 4

Adjusting Generalized RMA for model organisms

In this chapter we discuss how the group LASSO based generalized RMA framework developed in Chapter 3 can be easily modified for other applications for which locus effects require multiple predictors. Specifically, we will focus on the use of the haplotypes of a population's founders to identify loci that are associated with a phenotype within outbred crosses such as the Diversity Outbred (DO) or Heterogeneous Stock (HS) populations.

While association studies in model organisms allow for much stronger control of the genetics and experimental designs, they still have many of the issues we have discussed in the human setting. The most important to emphasize is that loci predictors are still highly confounded by LD, which results in highly confounded logP values when considering in standard single locus methods. Thus, there is also a great need for multiple locus methods that are less confounded by LD such as LLARRMA-haplo, described here.

4.1 Introduction

A number of experimental strategies have been proposed for association mapping of complex traits in model organisms. Many involve the use of highly recombinant populations derived from inbred lines. Examples of such populations are advanced intercross lines (AILs; proposed by Darvasi and Soller, 1995), where a pair of inbred founders are intercrossed for three or more generations, and heterogeneous stocks (HS; Demarest et al., 1999), where a number, usually eight, of inbred strains are intercrossed for many generations. The Diversity outbreed (DO; Svenson et al., 2012) population has been developed recently in mice based on collaborative cross (CC) founders, which resemble the HS in breeding structures. In theory, these strategies can achieve much higher-resolution mapping than that obtainable in standard inbred strain crosses. One such reason is that they accumulate a greater density of recombinations, allowing for a more fine mapping of the founders. An issue that presents in outbred populations is that the individuals in the population are related to some level, which often violates standard mapping techniques that may be applied to independent subjects. Further, because the markers used for genotyping will have fewer alleles than the number of haplotypes in the cross, individual markers typically do not unambiguously identify the underlying strain haplotype. In particular, unless all variants are genotyped, QTL will be missed by single-marker association analysis (Mott et al., 2000).

These highly recombinant structured experimental populations resemble those found in plant and animal breeding, where one common approach is to model the relatedness through variance components parameterized by the kinship matrix (Valdar et al., 2009). Specifically, the effects of a single locus are estimated simultaneously with one or more random intercept whose expected correlation structure is fixed given the pedigree (or realized genotypes) and that models the effects of overall genetic relatedness to account

for effects from the rest of the genome (Kennedy, Quinton and Vanarendonk, 1992; Jannink, Bink and Jansen, 2001; Zhao et al., 2007). Such approaches are applicable to HS, DO, and AIL populations, and reduce the false positive rate by reducing the effects and significance of loci that are predictive of family structure. This type of approach has been taken by two popular methods: Efficient Mixed-Model Association (EMMA) (Kang et al., 2008) and QTLRel (Cheng et al., 2011). EMMA was proposed as an efficient exact procedure that corrects for population structure and genetic relatedness in model organism association mapping during a period where it was not computationally efficient to use linear mixed effect models. While this was a great improvement, the EMMA algorithm was still computationally infeasible for large data sets because the variance components parameters are estimated for each marker (i.e., an exact solution). A new implementation of the algorithm called EMMAX (Kang et al., 2010) makes the simplifying assumption that because the effect of any given SNP on the trait is normally small, the variance parameters only need to be estimated once for the entire dataset, resulting in an approximate solution. This change sacrificed the exact solution calculation from EMMA for a feasible computation time. QTLRel (Cheng et al., 2011) is a more recent software which was developed to quickly perform genomewide scans, using a similar technique to EMMAX, with the potential of multiple random effects.

Although single locus polygenic based methods have been useful, as complex traits are affected by multiple functional loci, a multiple locus association method would be preferred (Ayers and Cordell, 2010). To identify the important loci within the multiple locus model, variable selection or regularization of the predictors is required (e.g., Sillanpää and Bhattacharjee, 2005; Hoggart et al., 2008; O'Hara and Sillanpää, 2009; Wu et al., 2009; Ayers and Cordell, 2010; Cho et al., 2010). The polygenic aspect of the model for both the distant (i.e., between populations) and close (i.e., within population) relatedness structures in the data can be addressed by a multiple

locus model, as the genetic relationships between the individuals can be captured by the markers themselves (e.g., Habier, Fernando and Dekkers, 2007). In Kärkkinen and Sillanpää (2012), they showed that multiple locus models that did not try to explicitly model polygenic effects worked well. Their observation of the redundancy in including additional polygenic components is in agreement with, for example, Calus and Veerkamp (2007) and Pikkuhookana and Sillanpää (2009).

Utz, Melchinger and Schön (2000) implemented a multiple locus resampling based procedure for detecting functional loci in GWAS, and showed in their simulations that the resampling was able to correct some biases and sampling errors in the model estimation. Schön et al. (2004) used a composite interval mapping regression approach (Haley and Knott, 1992) in combination with resampling of an multiple locus additive genetic model (as done in Utz, Melchinger and Schön (2000)) with loci selected by stepwise regression for the analysis of test cross progenies. They found that, even for moderate sample sizes, their procedure was able to obtain estimates with very low bias. They concluded that for traits regulated by a few QTL with large effects, for which phenotypic selection is expensive or hampered due to rare occurrence the resampling multiple locus approach of MAS (Utz, Melchinger and Schön, 2000) can be very useful.

Another resampling based multiple locus method called frequentist model averaging (FMA) was proposed in Hjort and Claeskens (2003). FMA examines each combination of predictors multiple locus models and averages over the models with weights to obtain parameter estimates. FMA can be implemented without much difficulty or protracted computations. One requirement of FMA is the specification of model weights. Several methods to define the weights have been proposed which include AIC weights (Buckland, Burnham and Augustin, 1997), weights based on minimizing a Mallows criterion (Hansen, 2007), and weights based on the Focused Information Criterion (Claeskens and Consentino, 2008). Williams and Christian (2006) showed that FMA estimates for

genetic effects in twins studies were more accurate than the standard estimates based on the criteria used for the model averaging weights. Schomaker, Wan and Heumann (2010) address the issue of missing data in the FMA framework. They proposed how one can incorporate imputation first and then preform FMA rather than attempt to incorporate complex weighting adjustments to criteria such as AIC which allow for missing data (e.g., the EM-based AIC developed in Claeskens and Consentino (2008)). They also propose a frequentist model selection (FMS) estimator which is a special case of FMA which focuses on the selected model rather than the estimated effects.

Here, we describe a new application of group based generalized RMA (see Chapter 3) which adds to the literature of multiple locus modeling for related individuals within a population. We introduce LLARRMA-haplo for interval association mapping with model predictors obtained from HAPPY and apply it to simulated HS populations.

4.2 Methods

Before discussing how we implement haplotype probability based predictors within the group LASSO generalized RMA framework developed for LLARRMA-dawg (see Chapter 3), we discuss in more detail the probability models obtained from HAPPY (Mott et al., 2000).

4.2.1 Diplotype Probability Models

Rather than the traditional observed marker data, we are interested in modeling the subjects haplotypes in the intervals between observed markers. In the context of multiple founder crosses, we can use the detailed founder haplotype information to identify the state of each subject in the interval. In brief, haplotype descent along each subject's genome can be inferred by the haplotype reconstruction method HAPPY (Mott et al., 2000), which applies a hidden Markov model simultaneously to the genotypes of the founder strains and the n subjects. For each subject HAPPY produces a vector $\boldsymbol{g}_i(m)$

for each interval i (i.e., between adjacent pairs of observed markers), which contains the descent information from the founders at marker m based on either an additive or full effect model which are described in detail below.

Before describing the exact form of $\boldsymbol{g}_i(m)$, let us consider a cross with J founders. For each locus, the subject within the cross will have two haplotypes present, one on each copy of the chromosome where the locus resides. For each individual, HAPPY will provide us with a $J \times J$ matrix $\mathbf{P}$, where p_{ij} is the probability that the first haplotype is from founder i and the second haplotype is from founder j. We summarize $\mathbf{P}$ as $\boldsymbol{g}(m)$ based on one of the selected model described below.

Additive Model

The additive haplotype model describes the locus based on the expected number of each founders haplotype present at each given locus. For a J founder cross, the additive version of $\boldsymbol{g}(m)$ is a J-vector and which sums to 2. The exact definition of the additive locus predictor for subject i at locus m is given by

$$\boldsymbol{g}_i^a(m) = \mathbf{1}^{\mathrm{T}}(\mathbf{P} + \mathbf{P}^{\mathrm{T}}), \tag{4.1}$$

where $g_j^a = E(\text{number of haplotype } j)$ and $\mathbf{1}^{\mathrm{T}}\mathbf{g}^a = 2$

Full Model

The full diplotype model describes the locus based on the probability of each unique founder haplotype pair (or diplotype) at each given locus. For a J founder cross, the full model version of $\boldsymbol{g}(m)$ is a $J(J-1)/2$ length probability vector. The exact definition of the full model locus predictor for subject i at locus m is given by

$$\boldsymbol{g}_i^f(m) = \text{vech}\left(\mathbf{P} + \mathbf{P}^{\mathrm{T}} - \text{diag}(\text{vecdiag}(\mathbf{P}))\right), \tag{4.2}$$

where vech() returns the upper triangle matrix, including the diagonal, as a vector, vecdiag() returns the diagonal as a vector, and $\mathbf{1}^{\mathrm{T}}\mathbf{g}^{f}$ =1.

4.2.2 LLARRMA-haplo Framework

We start by considering the use of standard linear regression to estimate the effects of marker intervals m in the set of considered markers $\mathcal{M}$ containing M markers on a quantitative outcome from n individuals. We then describe statistical approaches to identify a subset of m_q intervals that are truly influential. Here we define a "true signal" to be a marker interval that contains an underlying causal variant, a "background" interval to be a interval that is not a true signal, and an optimal analysis as one that distinguishes true signals from background intervals. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be an n-vector of quantitative responses. Let $\mathbf{g}_i(m)$ be the haplotype predictors as described for the additive (see Eq. 4.1) or full (see Eq. 4.2) model for loci $m \in \mathcal{M}$.

We model the quantitative phenotype of individual i by a linear regression of the $8M$ (or $36M$) predictors for the additive (or full) haplotype effects model as

$$y_i = \mu + \sum_{m \in \mathcal{M}} \boldsymbol{\beta}_m \mathbf{g}_i(m) + \epsilon_i, \tag{4.3}$$

where μ is the intercept, $\boldsymbol{\beta}_m$ are the effects of the haplotype predictors for locus m, $\mathbf{g}_i(m)$ are the haplotype predictors for individual i as described for the additive or full models for locus m and $\epsilon_i \sim N(0, \sigma)$ is the error term.

In this setting, each locus m has a set of predictors, $\mathbf{g}(m)$, which are required to model the effects at the locus. This falls naturally into the grouped generalized RMA framework discussed in Chapter 3. Specifically, we can apply generalized RMA to our model with variable selection performed by the group LASSO defined by

$$\hat{\boldsymbol{\beta}}_{\mathcal{M}}(\lambda) = \underset{\mu, \boldsymbol{\beta}_m; m \in \mathcal{M}}{\operatorname{argmin}} \left\{ -\ell(\mu, \boldsymbol{\beta}_m; m \in \mathcal{M}) + \lambda \sum_{m \in \mathcal{M}} \|\boldsymbol{\beta}_m\|_2 \right\}, \tag{4.4}$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ is a vector containing each $\boldsymbol{\beta}_m$ for $m \in \mathcal{M}$, $\ell(\cdot)$ is the log likelihood and $\|\cdot\|_2$ is the l_2 norm. We note the abuse of notation in that the argmin returns μ and $\boldsymbol{\beta}_m; m \in \mathcal{M}$ and we disregard μ as we are only interested in evaluating which loci are included.

Applying the generalized RMA framework to our model with variable selection done by the group LASSO (as given by Eq. 4.4) for K resamples yields the RMIP estimate for locus m

$$\widehat{\text{RMIP}}_m = \frac{1}{K}\sum_{k=1}^{K} \hat{\gamma}^{(k)}(\boldsymbol{\beta}_m)\,, \tag{4.5}$$

where

$$\hat{\gamma}^{(k)}(\boldsymbol{\beta}_m) = \begin{cases} 1 & \text{if } \boldsymbol{\beta}_m \neq \mathbf{0} \text{ for resample } k \\ 0 & \text{if } \boldsymbol{\beta}_m = \mathbf{0} \text{ for resample } k \end{cases}. \tag{4.6}$$

4.2.3 Completing Methods

The standard methods which are used to analyze the populations of interest are all single locus based. A summary of the methods are given below.

Naive Single Locus

The simplest approach to analyze the data would be using the naive single locus model which we have previously applied to populations where the individuals are assumed to be independent (see Chapters 2 and 3). This simple linear model (or generalized linear model) ignores the information about the related subjects, which results in the errors not being independent.

Efficient Mixed-Model Association (EMMA)

EMMA (Kang et al., 2008) is a mixed effects modeling software which has been adapted for many genetic association modeling frameworks. The simplest version of EMMA simply models the effect of a single predictor (e.g. an additive SNP effect) with a

random effect such as the kinship relationship for related individuals. In EMMA's most recent rebranding, EMMAX (Kang et al., 2010) it is described in supplementary materials how you can use the REMLE estimates for the variance components of the random effects to represent the random effects model as a generalized least squares (GLS) model. Once in the GLS form, it is simple to transform the model into a simple linear model to compare models which have more than one additional predictor in the alternative model (e.g. additive or full haplotype models). We use this convention to evaluate EMMA on our simulations. We note that the resulting logPs obtained are only approximate as the random effect variance is not re-estimated under the alternative model and will not be at its maximum for loci with true effects.

QTLrel

QTLrel (Cheng et al., 2011) is another single locus mixed effects model designed for populations which have related individuals. QTLrel uses the pedigree to identify the additive and dominant genetic random effects structures, but as we do not have full pedigree information we use the realized kinship matrix (as done in EMMA) to infer the structure of the additive random effect's structure. QTLrel can be used for genotype scans and also scans on the additive or full model founder probabilities. For all analyses, QTLrel uses the same type of transformed linear model approximation we have discussed for EMMA under haplotype probabilities. QTLrel has the advantage that it is capable of incorporating additional random effects (e.g. batch or cage effects) which we have not simulated in our models.

4.3 Simulation Framework

4.3.1 Heterogeneous Stock: Population A

To test our method on a complex population with ambiguous descent, we first examined the HS population simulations from Valdar et al. (2009). Here, they simulated

100 populations of 500 F53 heterogeneous stock individuals derived from eight inbred lines. Modeling a minimal two-chromosome genome, they used marker genotypes from the HS study of Valdar et al. (2006). This comprised 870 markers spanning 98.6 cM on chromosome 1 and 759 markers spanning 103.7 cM on chromosome 2. All markers were diallelic with minor alleles distributed variously among the eight founder strains (see http://gscan.well.ox.ac.uk/ for more information). They simulated two diallelic QTLs on chr 1 and, to allow the simulation to focus on discrimination of signals rather than on power, they where positioned in marker-dense regions at 29 and 68 cM with additive effects each accounting for 10% of the phenotypic variance. The QTL acted in the same direction in the founders, had alleles split equally among the eight inbreds, but had strain distribution patterns that differed from those of their flanking markers. Each population was generated by a single funnel of four two-way crosses, two four-way crosses, and one eight-way cross, giving rise to a mating population of 100 individuals that was then circular-mated for 50 generations, with the mating pairs in the penultimate generation bred to produce 10 offspring each (see Valdar et al. (2006) and references therein).

Performance was assessed as for the advanced intercross trials by defining genome segments. Because the HS are more recombinant, segments were 6 cM wide and defined such that each QTL sat at a segment midpoint.

4.3.2 Heterogeneous Stock: Population B

Our second simulation setup follows the same basic outline of population A, but with the addition of multiple unobserved tiny effects on chromosome 2 which formulate a stronger polygenic effect in the population. Specifically, we simulate 100 populations of 500 F53 heterogeneous stock individuals derived from eight inbred lines. We start with the same two additive diallelic QTLs on chr 1 positioned in marker-dense regions at 29 and 68 cM, each accounting for 10% of the phenotypic variation, and add an

additional 50 unobserved minor effect QTLs evenly spaced across chr 2, accounting for approximately 50% of the phenotypic variability. The addition of the 50 tiny effect QTLs adds a much stronger heritable effect to the population.

Performance within this population was assessed based on the performance of the method to detect the two main effects on chromosome 1. We use the same 6 cM segment approach used on population A.

4.4 Simulation Results

4.4.1 Results from 100 simulations in HS population A

Below we discuss the results from 100 simulations of an HS population with 2 chromosomes where each simulations contains two true loci on chromosome 1.

Additive model results

Figure 4.1 displays the ROC curves for the additive haplotype model based on 100 simulations from HS population A. We observe that the naive single locus model is clearly out performed by all other methods. The performance of LLARRMA-halpo, EMMA, and QTLrel appear to be indistinguishable based on their performances with the additive model.

Full Model results

Figure 4.2 displays the ROC curves for the full diplotype model based on 100 simulations from HS population A. We observe that the naive single locus model (AUC of 0.785) is still clearly out performed by all other methods. The performance of LLARRMA-halpo (AUC of 0.905) now clearly performs better than both EMMA (AUC of 0.8729) and QTLrel (AUC of 0.8735) based on their performances with the full model.

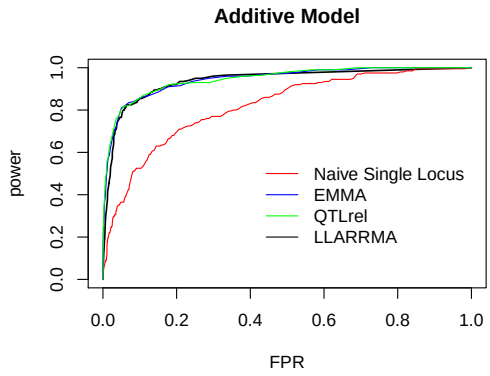

Figure 4.1: ROC curves for the additive model based on 100 simulations on HS population A. We observe a clear advantage to methods with either multiple locus modeling (LLARRMA-haplo) or mixed effect models (EMMA and QTLrel).

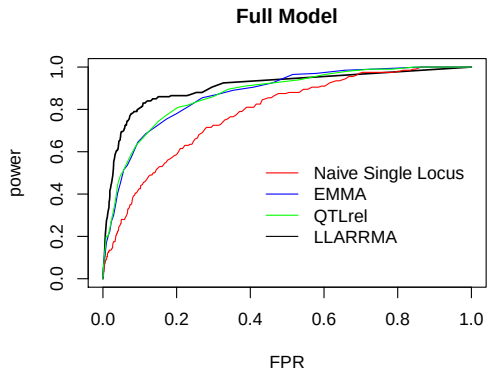

Figure 4.2: ROC curves for the full model based on 100 simulations on HS population A. With the increase to the full model, we observe a advantage to LLARRMA-haplo over mixed effect models (EMMA and QTLrel).

4.4.2 Results from 100 simulations in HS population B

Below we discuss the results from 100 simulations of an HS population with 2 chromosomes. Each simulations contains two true loci on chromosome 1 and 50 unobserved loci on chromosome 2 which provide a family based effect on the phenotype. This model more closely resembles the setting of fine mapping studies of complex phenotypes than population A.

Additive model results

Figure 4.3 displays the ROC curves for the additive haplotype model based on 100 simulations from HS population B. We observe that the naive single locus model (AUC of 0.621) is clearly out performed by all other methods. The performance of EMMA (AUC of 0.856), and QTLrel (AUC of 0.859) appear to be indistinguishable based on their performances with the additive model. The performance of LLARRMA-halpo (AUC of 0.786) is slightly worse than that of EMMA or QTLrel based on the additive model.

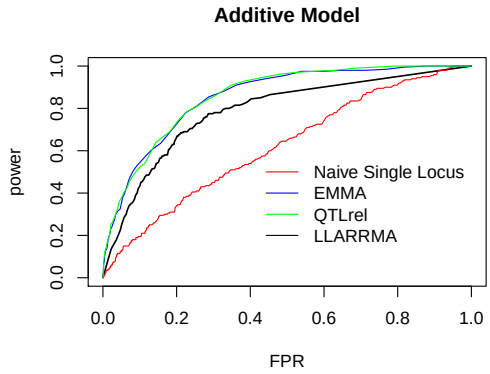

Figure 4.3: ROC curves for the additive model based on 100 simulations on HS population B. We observe a clear advantage to methods with either multiple locus modeling (LLARRMA-haplo) or mixed effect models (EMMA and QTLrel) with LLARRMA-haplo performing slightly worse.

Full Model results

Figure 4.4 displays the ROC curves for the full diplotype model based on 100 simulations from HS population A. We observe that the naive single locus model (AUC of 0.606) is still clearly out performed by all other methods. The performance of LLARRMA-halpo (AUC of 0.772) now clearly performs better than both EMMA (AUC of 0.693) and QTLrel (AUC of 0.681) based on their performances with the full model.

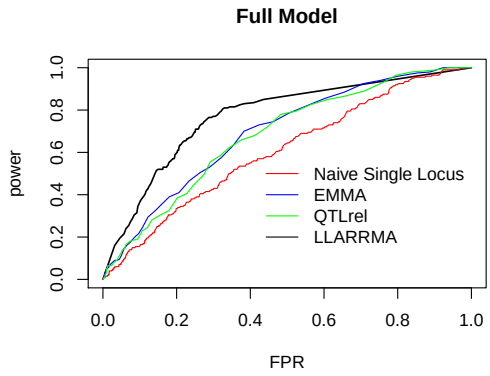

Figure 4.4: ROC curves for the full model based on 100 simulations on HS population A. With the increase to the full model, we observe a advantage to LLARRMA-haplo over mixed effect models (EMMA and QTLrel).

4.5 Discussion

We present LLARRMA-haplo, a new application of our previous group generalized RMA framework for characterizing frequentist variability in model selection, and apply them to a well suited problem, discriminating true signals from false signals in highly recombinant populations derived from multiple founder crosses. Such populations can be used to map loci far more accurately than possible with standard intercrosses. However, the varying degree of relatedness that exists between individuals complicates the analysis. The recent consensus in the animal breeding community, and elsewhere, suggests explicit model selection even in the absence of kinship-like modeling is preferable to models including a polygenic effect. We add to this literature by exploring the multiple locus approach of LLARRMA-haplo against some standard methods such as EMMA and QTLrel.

Based on performance in our simulations, LLARRMA-haplo appears to be competitive with popular polygenic effect methods based on observed kinships such as EMMA and QTLrel. When we consider only an additive model, LLARRMA-haplo appears to be competitive only with the polygenic effect models, whereas when we look at the more

complex full model, LLARRMA-haplo outperforms the polygenic models in each of our simulations. In each simulation, the underlying QTL effects are additive, indicating that the additive model is sufficient for modeling and the full model, which can capture additive and dominance effects, includes additional unnecessary predictors. Comparison of each methods additive and full model performance shows that LLARRMA-haplo is not negatively affected by the more complex model when the simpler model is appropriate, whereas the single locus models have a large drop in performance when switching from the additive to full model. This indicates a potential advantage to LLARRMA-haplo in situations where a more complex model may be needed to capture the underlying effects.

Although we take the approach of multiple locus modeling over polygenic random effects, when our model selection procedure fails to include the entire set of true loci in our model we are not able to fulfill the full potential of multiple locus modeling. While we take a discovery-based permutation approach to selecting λ in the group LASSO, with the aim of removing noise variables only, we will fail to include some true predictors. If we can assume that it is a negligible amount of true signals which will be missed, then our approach is well justified. On the other hand, if we are unable to capture the majority of the true signals, then our approach may fall short of methods that account for the related individuals in other ways. Our HS population B simulation is an example of a large number of minor effects which may appear to be noise and fail to be included by our permutation selection of λ. In this setting we saw that the multiple locus modeling approach underperformed when compared to polygenic models.

To address the issue of missed signals in the multiple locus modeling approach, we have a few possible directions which may lead to improved performance. If we assume that the setting of HS population B is a common setting, we may want to include a random effect for a polygenic effect in the model to capture the effects of signals

we do not include. This would be very computationally intensive to do. If we took this approach, we may want to change the structure of the polygenic effect to only account for variables which have been excluded from the model which would greatly increase the computational complexity. A similar approach which would have much less computational complexity would be to run LLARRMA-haplo on the residuals from a polygenic model with no loci included. Although this may help address this issue, it has a potential to regress out real effects in the polygenic model which we would not be able to recover. A third approach may be to include additional non-penalized fixed effects in the model to account for additional structure that is missed by excluded variables. A simple way to do this may be to include principle components in the model as unpenalized covariates.

Although there are still some minor aspects to LLARRMA-haplo which need to be addressed, it shows great potential as a method for association mapping in highly recombinant populations derived from multiple founder crosses. We have seen that the model averaging approach of LLARRMA-haplo is able to handle more complex models than the underlying model without the loss of power which is present for most standard single locus models. This allows for the use of models which can detect both additive and dominance effects without the loss of power when the model is only additive, and may potentially lead to increased power when non-additive effects are present.

Chapter 5

Applications of Generalized RMA

This chapter discusses some of the real data sets that have been analyzed by forms of LLARRMA. We focus on the application of LLARRMA and LLARRMA-dawg in select human GWAS hit regions, and the application of LLARRMA-haplo to the Diversity Outbreed (DO) and Heterogenous Stocks (HS) populations.

5.1 Human GWAS data

We discuss select results from two data cohorts whose purpose was to identify cardiovascular disease (CVD) risk factors. Below we discuss these data sets and the genotyping arrays used to obtain the data.

5.1.1 Atherosclerosis Risk in Communities Study (ARIC)

ARIC is a longitudinal cohort study of atherosclerosis. It is a population-based sample of 15,792 men and women aged 45 to 64 years who were recruited from 4 US communities (Forsyth County NC, Jackson MS, suburban Minneapolis MN, and Washington County MD) between the years of 1987 and 1989. Participants of ARIC received an initial extensive examination for medical, social, and demographic data. Participants were reexamined every three years, with the last between the years of 1996 and 1998.

5.1.2 Multi-ethnic Study of Atherosclerosis (MESA)

MESA is a cohort study of the characteristics of subclinical cardiovascular disease and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA includes a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Demographically, MESA consists of 2,622 whites (39%), 1,893 African-Americans (28%), 1,496 Hispanics (22%) and 803 (12%) of Asian (primarily Chinese) descent. Participants were recruited from six field centers across the United States (Winston-Salem, NC; St. Paul, MN; Chicago, IL; Los Angeles, CA; New York, NY; Baltimore, MD).

5.1.3 IBC genotyping

The IBC SNP array is described in detail in Keating et al. (2008). The IBC SNP array includes 49,320 SNPs selected across 2,000 candidate loci for CVD. The array includes SNPs that capture patterns of genetic variation in both European- and African-descent populations. Genotyping for the CARe cohorts (which include ARIC and MESA) was performed at the Broad Institute (Cambridge, MA). Criteria for DNA sample exclusion based on genotype data included sex mismatch, discordance among duplicate samples, or sample call rate $< 95\%$. For each set of duplicates or monozygotic twins, data from the sample with the highest genotyping call rate were retained. SNPs were excluded when monomorphic, the call rate was $< 95\%$, or HWE was $p < 10^5$ in EAs. Given the genetic admixture in African Americans, there was no HWE filter used for these samples. After these exclusions were applied, data remained on 47,539 SNPs.

5.1.4 Zoom Locus plots

LocusZoom (Pruim et al., 2010) is a common tool used to plot human GWAS hit region results. LocusZoom is widely used for plotting hit region results as it is able to display LD information with reference to the top hit and also displays many of the

known genes within the region. The LD information allows one to have an idea of how much potential confounding from LD is present, which is useful when attempting to determine the number of underlying loci in the data. The additional gene information given by LocusZoom is also of great use in practice for considering potential functional effects SNPs may have based on their location with respect to a gene. We adopt the LocusZoom plot for our LLARRMA human hit region analyses for an easy comparison with single locus scans using LocusZoom.

5.1.5 Cardiovascular Disease Risk analyses

CVD is a highly complex trait that is influenced by many genetic and environmental variables, which makes it very difficult to study directly. To better understand the genetic components of CVD, it is common to perform separate analyses for different genetic CVD risk factors. In the following sections we present select results from LLARRMA and LLARRMA-dawg on hit regions for CVD risk factors that where found interesting by the researchers that the data belong (Ethan Lange and Leslie Lange, UNC Genetics department).

ARIC: African American GWAS data

We highlight the analyses of two phenotypes from the ARIC African American (AA) GWAS data, C-reactive protein (CRP) and factor VII (FVII). For the CRP GWAS data, we examine a hit region on chromosome 1 near the CRP gene which is highly confounded based on single locus regression, displaying 8 SNPs which are nearly indistinguishable based on marginal logPs. For the FVII GWAS we examine a less complex hit region from chromosome 13 near the FVII gene.

Figure 5.1 displays the single locus regression and LLARRMA outputs from the CRP hit region. We observe that when using the standard single locus scan (top) we have 8 SNPs with logP values are all highly significant. Four of the SNPs have nearly identical logP and the remaining four are just slightly less significant. Each of these

SNPs are in very high LD making it very difficult to make any inference about which may be a true signal or even how many true signals may be present. Examining the LLARRMA output (bottom), we observe that the most significant SNP based on SL (rs7531832) has a very low RMIP while SNP rs16827466 which was just slightly less significant based on SL logP presents with an RMIP of 0.9, giving strong evidence of the SNP being a true signal. The LLARRMA output also provides reasonably strong evidence (RMIP greater than 0.6) of three additional SNPs, potentially suggesting that CRP has multiple true signals within this region. It has been debated that CRP has multiple causal SNPs within this region, but it has been hard to justify this claim based on single locus methods. LLARRMA provides some justification to this hypothesis.

Figure 5.2 displays the SL and LLARRMA outputs from the FVII hit region on chromosome 13. The SL logP (top) shows a highly significant SNP (rs1755685) and a few SNPs in moderate to low LD with the top SNP also presenting with significant logPs. With the large drop in significance between the SNPs, it is likely that one would expect that there may only be one true signal in the region based on the SL analysis. LLARRMA (bottom) gives a much sparser output within the region, which is consistent with the SL top SNP observing a RMIP of 1. LLARRMA provides evidence of a potential importance of SNP rs547138 (RMIP of 0.8) that was not easily observed in the SL logPs. This example shows that even in less complex regions (where LLARRMA does not have a large advantage in performance), the observed output is much simpler to interpret.

ARIC: European American IBC genotypes

Here we emphasize a hit region for which LLARRMA and LLARRMA-dawg have significant differences in the RMIPs, potentially indicating the presence of a non-additive effect. Figure 5.3 displays the LLARRMA and LLARRMA-dawg outputs for the hit region for HDL located on chromosome 8 within in the ARIC data set. We observe

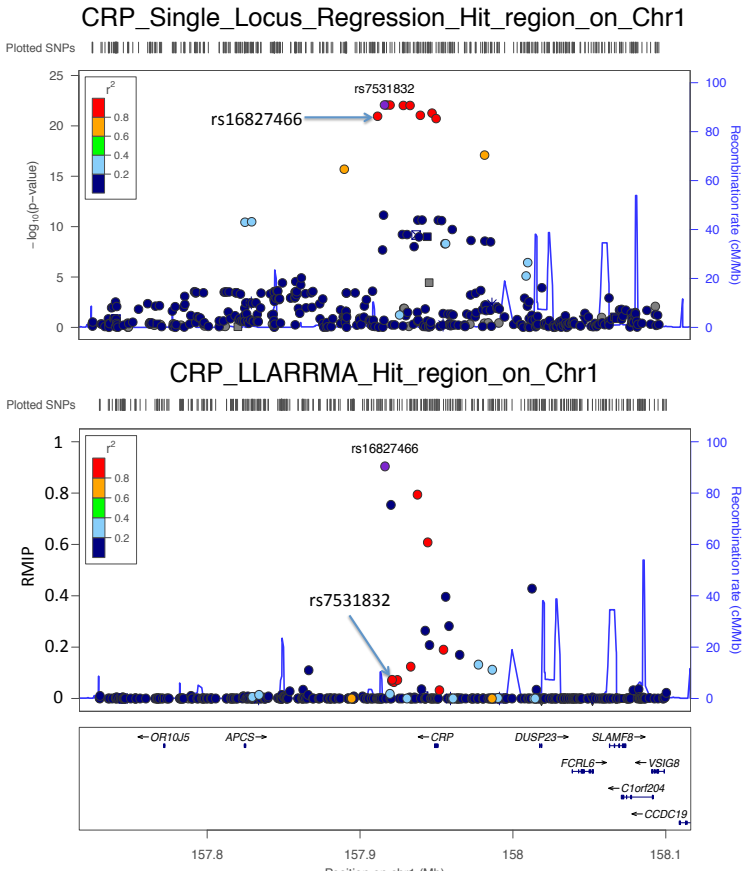

Figure 5.1: Single locus regression and LLARRMA outputs for ARIC African American GWAS data hit region on chromosome 1 for CRP. We observe a large set of significant SNPs in the single locus approach are hard to distinguish between, while the LLARRMA output has a smaller set of defined SNPs with high RMIPs.

that by allowing for a non-additive model, LLARRMA-dawg finds additional SNPs that where not given high priority based on the additive only model used by LLARRMA. We observe agreement between the SNPs with high RMIPs from LLARRMA when examining the RMIPs from LLARRMA-dawg. We observe that LLARRMA-dawg found two additional SNPs of potential importance (rs11570891 and rs894210). These two SNPs received very low RMIPs based on an additive only model, but when fitting the general model which allows for dominance effects we find that they are important. This

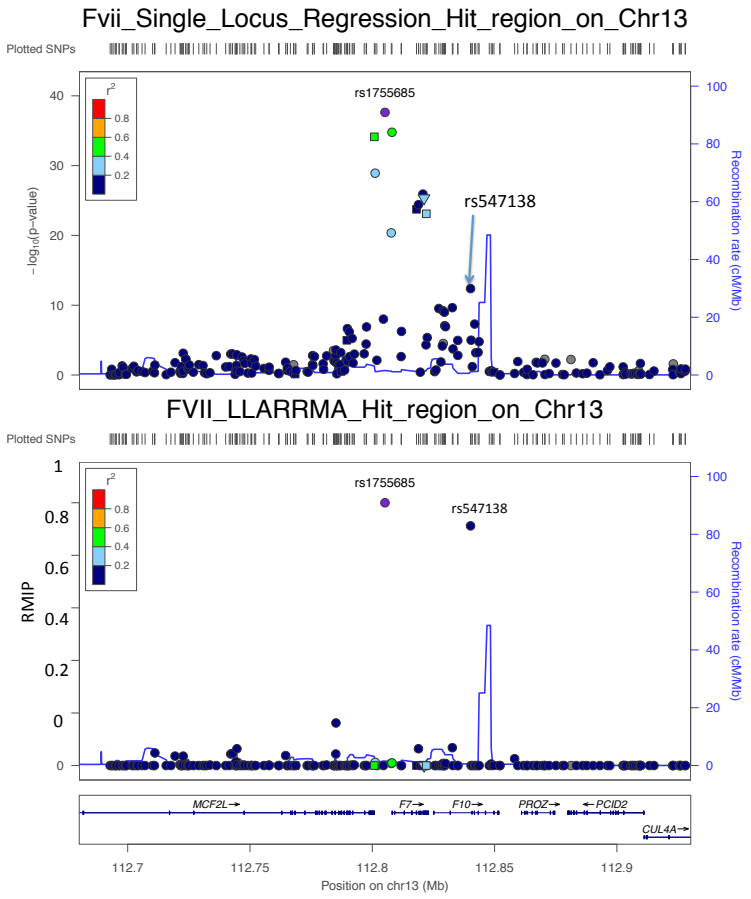

Figure 5.2: Single locus regression and LLARRMA outputs for ARIC African American GWAS data hit region on chromosome 13 for factor 7 levels. We observer both single locus and LLARRMA selecting the same top SNP, but LLARRMA highlights the importance of a second SNP which was not as obvious from the single locus scan.

potentially indicates that these SNPs may be true signals that have non-additive effects which would be missed based on additive effect methods.

MESA: European American IBC genotypes

The Analysis of the MESA European American data set lead to another interesting result for the CRP phenotype. We examine the same region as we found on chromosome 1 in the ARIC African American GWAS. We observe that, for both the standard single locus and LLARRMA, it appears that there is a different set of SNPs driving the CRP

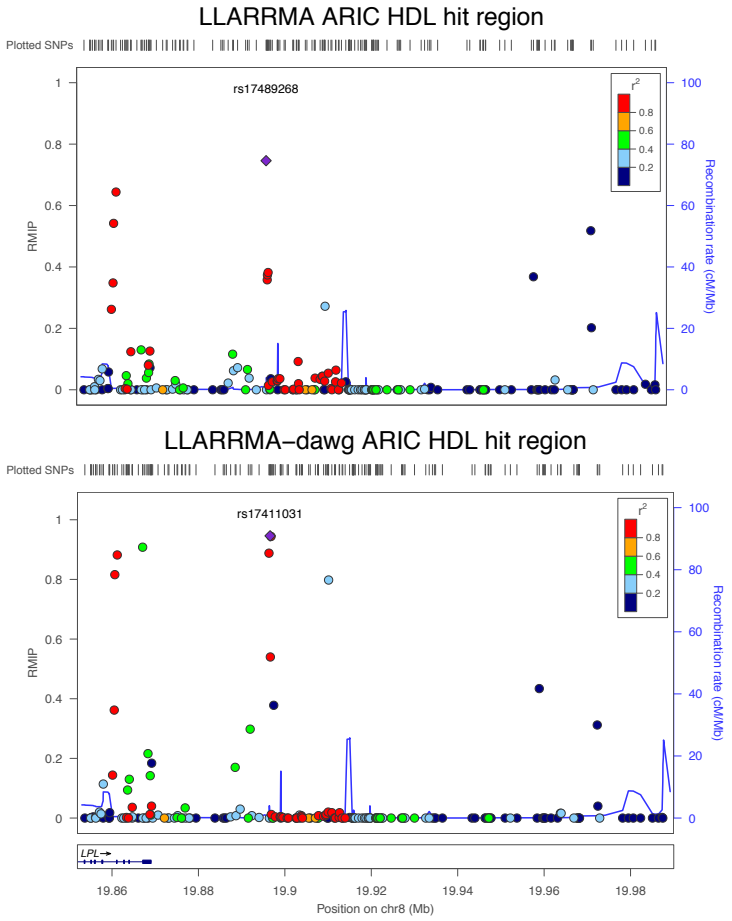

Figure 5.3: LLARRMA and LLARRMA-dawg outputs for ARIC European Americans hit region for HDL on chr 8.

phenotype within this region, which is consistent with the assumption that African Americans and European Americans may have different effects for CVD related traits.

Figure 5.4 displays the MESA hit region on chromosome 1 for CRP. Examining the single locus scan (top) we observer 4 SNPs with logPs of about 20, three which are within the CRP gene and one which is found down stream of the gene. If we were to base a conditional regression analysis on the single locus we would select the top SNP labeled in the figure. When we examine the LLARRMA output, we observe that the 3 SNPs SL had found within the CRP gene where never selected in any resample, while

the forth SNP (rs12567054) which was up stream was selected in every resample. We also see evidence of a second signal (rs2794515) which has a RMIP of above 0.9.

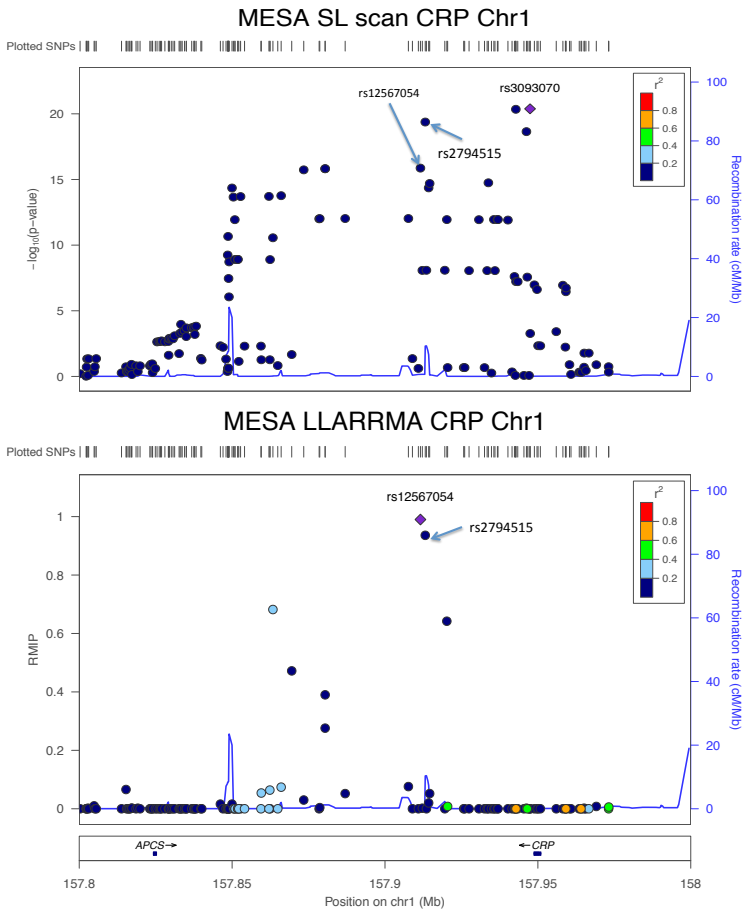

Figure 5.4: Single locus regression and LLARRMA outputs for MESA European Americans hit region for CRP on chr 1.

One possible explanation of these findings is that LLARRMA has selected regulatory SNPs for the CRP gene which would directly lead to changes in the levels of CRP. While the SNP selected based on SL may be important in the gene, regulatory eliminates may have a much larger impact on CRP, which is why they were preferred by LLARRMA. It is also possible that selected SNPs fall into another gene in the CRP pathway which shows higher signal in this data. Follow up experiments would be required to examine what is driving the CRP levels in this region to try and validate the hypothesis that

LLARRMA has selected regulatory SNPs for the CRP gene, or another gene involved in the CRP pathway.

5.2 Model Organism data

To show the usefulness of the LLARRMA-haplo framework described in Chapter 4, we present preliminary analyses on the HS population.

5.2.1 Heterogeneous Stock (HS) Mice

The HS data is described in full detail in Valdar et al. (2006). In brief summary, genotypes for 13,459 SNPs on 1,904 fully phenotyped mice and 298 parents were obtained with an accuracy of 49.9%. The heterogeneous stock tested is comprised of a complex pedigree, resulting in linkage disequilibrium (LD) between pairs of markers which is complex (making the population well suited for application of LLARRMA-haplo). A high-throughput phenotyping protocol was used to collect multiple phenotypes as described in Solberg et al. (2006) (and is available online at http://gscan.well.ox.ac.uk/).

Analysis of Mean Adrenal Weight

We analyzed the HS data with the Mean Adrenal Weight (MAW) as a phenotype. We selected this phenotype for a proof of principle analysis for LLARRAM-haplo based on its complex nature that was observed by more traditional analyses such as single locus scans (see http://gscan.well.ox.ac.uk/ for details). Figure 5.5 displays (top) the single locus scan and (bottom) the LLARRMA-haplo RMIP output based on the full diplotype model. We observe that many of the locations with high RMIPs match with SL peaks, many of which are less significant in comparison to other SL peaks which were not selected by LLARRMA-haplo. Note that the most significant peak based on single locus scans (located on chromosome 15) corresponds with the findings of LLARRMA-haplo.

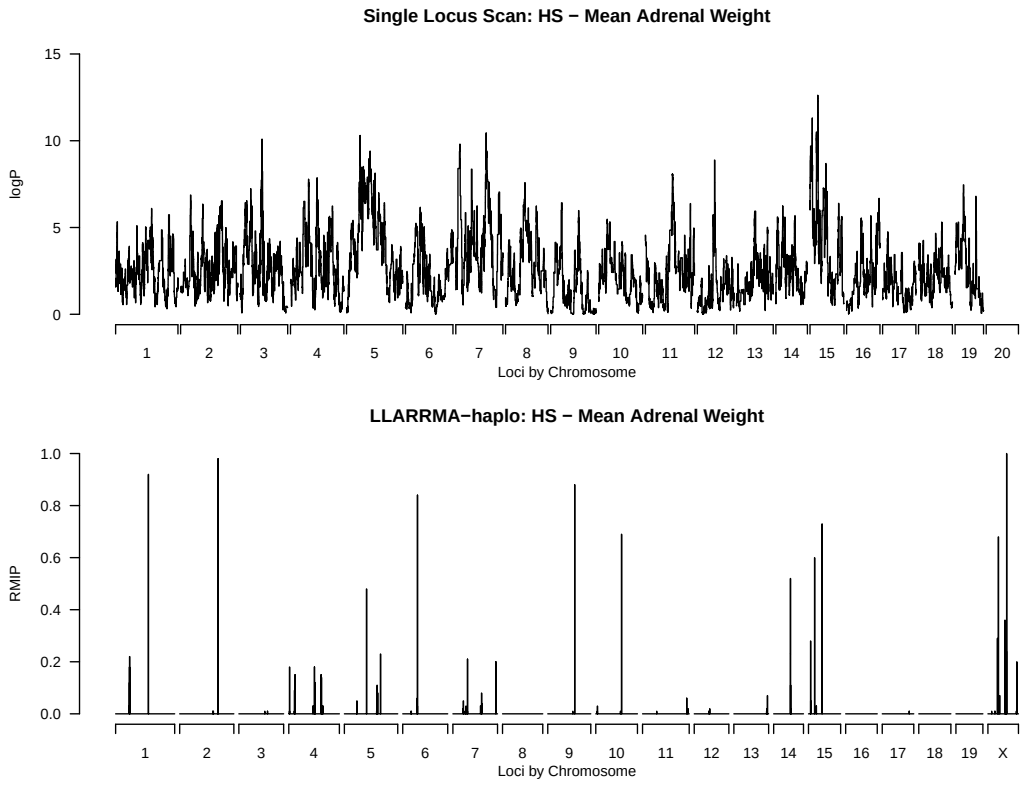

Figure 5.5: LLARRMA-haplo output for HS mice for Mean Adrenal Weight.

Chapter 6

Higher dimensional RMA - 2D-RMIPs

In this chapter we will discuss a technique that can be preformed using the information recorded when applying a RMA procedure. We focus on a higher order summary of the selected variables. Specifically, we examine the pair-wise resample model inclusion probabilities, or a 2D-RMIP, and examine additional information that can be observed from them.

6.1 Response relevant predictor relationships

One of the first analyses that is preformed when examining a data set is to examine the correlation structure between the variables. In a GWAS, one would do this by examining the LD, often measured as the square of the correlation. Although this information can be very useful, when the goal of your analysis is to identify the true variables in the model from a large set of potential candidates, simply knowing the correlation between all the variables may not provide any useful information about the underling problem of interest.

When simply examining the correlations between each of the variables, many of the pairwise correlations provide no 'relevant' information about the variables in the underlying model. When considering a sparse model, as is most often the case in

genetic data, it can be the case which the majority of the information is not relevant to the underlying model. Specifically, knowing that two variables are highly related by correlation, does not imply that they are both related to the response. Likewise, knowing two variables are close to independent does not mean that they do not work together in the model for the response.

As we are interested in identifying information about the relationship between variables with respect to the response, we propose to examine the 2D-RMIP as a measure of information between the variables relevant to the response. Variables which are related to the response would be expected to have high 2D-RMIPs (i.e. true predictors are often jointly included in the selected model), while variables which have no relationship within the model would be expected to have low 2D-RMIPS (i.e. background variables are not often jointly selected). We define the 2D-RMIP between variables i and j as

$$2\text{D} - \text{RMIP}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \Gamma_{ik} * \Gamma_{jk} \,, \tag{6.1}$$

where $\mathbf{\Gamma}^{\mathrm{T}} = [\hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\gamma}}^{(2)}, \cdots, \hat{\boldsymbol{\gamma}}^{(K)}]$ is the matrix of selected models from the RMA procedure.

We propose to use the 2D-RMIP as a measure of response relevant information between variables. We illustrate how this information may be useful with the following toy example.

6.1.1 Motivating toy example

We consider a simple simulation based on the hapgen2 data sets used in Chapter 3. We simulated a less sparse model, including 10 true SNPs in the model, with a higher signal to noise ratio to than used for the LLARRMA-dawg simulations. With the response generated, LLARRMA was run on the data. For visualization of the 2D-RMIP, we

restrict the SNP set to SNPs that have an RMIP of at least 0.25, as these are the only variables with supporting evidence to be included in the model based on LLARRMA.

Figure 6.1 displays the both the LD, measured as r^2, (left) of the selected SNPs and the 2D-RMIPs (right) for the selected set of SNPs. The true SNPs have been highlighted with dashed red lines. Examining the LD of the selected SNPs, it appears that most of the SNPs are not related. When we examine the 2D-RMIPs, we observe a structure that does well at highlighting the SNPs which are related in the simulated model. We observe that the variables which are related based on their effects on the phenotype appear to not be related based on LD, while the 2D-RMIP does a nice job of highlighting their relationships relevant to the response.

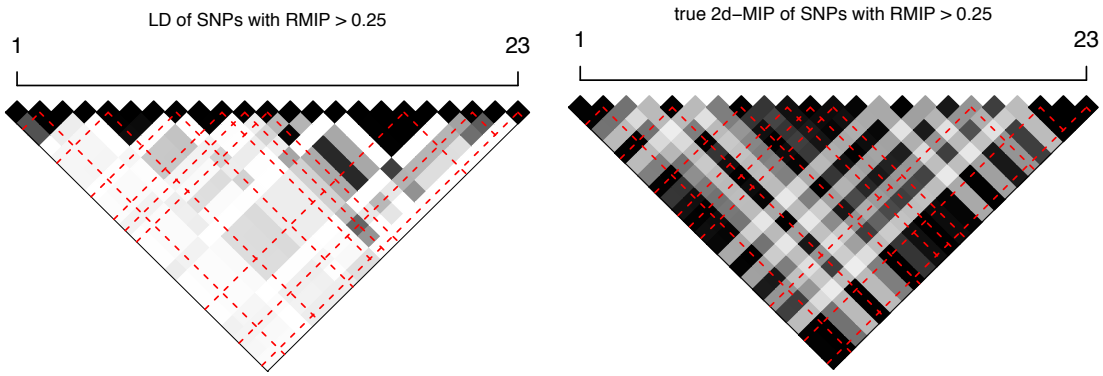

Figure 6.1: (Left) displays the LD between SNPs that had RMIPs of at least 0.25. (Right) displays the 2D-RMIP of the same variables. Red lines indicate true SNPs in the model. We observe that the 2D-RMIP does well identifying pairs of variables which have true effects with the response.

6.1.2 Real Data application

To illustrate the use of the 2D-RMIP, we apply LLARRMA to the The Cancer Genome Atlas Network (TCGA) breast cancer data described in TCGA (2012). While some SNP data was available for this data set, we explore the use of LLARRMA with genome-wide gene-expression (GE) variables. We explore how GE values can be used to compare

tumor vs normal tissue, and also to distinguish between Luminal A and Luminal B cancer subtypes.

The TCGA Breast Cancer data

Tumor and germline DNA samples were obtained from 825 patients. Patients were assayed on different subsets of platforms. 466 tumors from 463 patients had data available on five platforms including Agilent mRNA expression microarrays (n = 547), Illumina Infinium DNA methylation chips (n = 802), Affymetrix 6.0 single nucleotide polymorphism (SNP) arrays (n = 773), miRNA sequencing (n = 697), and whole-exome sequencing (n = 507); in addition, 348 of the 466 samples also had reverse-phase protein array (RPPA) data (n = 403).

Rsem GE values were normalized to the upper quartile of the total counts. The data was then log2 transformed and genes were median centered. Genes were filtered such that at least 70% of samples had a value (17,007 genes), and then imputed.

Analysis: Cancer vs. Adjacent normal

We start with a simple normal breast tissue vs breast cancer tumor analysis where we applied LLARRMA to the GE values for samples. Figure 6.2 displays (top) the LLARRMA output with a gray dashed line indicating the RMIP threshold for 2D-RMIP comparisons and (bottom left) the r^2 values and (bottom right) 2D-RMIPs for variables with RMIPs greater than 0.25. When comparing the correlation of the GE values we observe that most of the highlighted variables from LLARRMA, those with RMIPs above 0.25, have low correlations and do not seem to be highly related. When we examine the 2D-RMIP plot, we observe that many of the moderately correlated genes appear to be highly related with respect to the cancer/healthy tissue phenotype.

Analysis: Luminal A vs. Luminal B

We follow up the simple cancer/healthy tissue analysis with a subtype specific analysis. Specifically, we examine the two 'Luminal' breast cancer subtypes, LuminalA

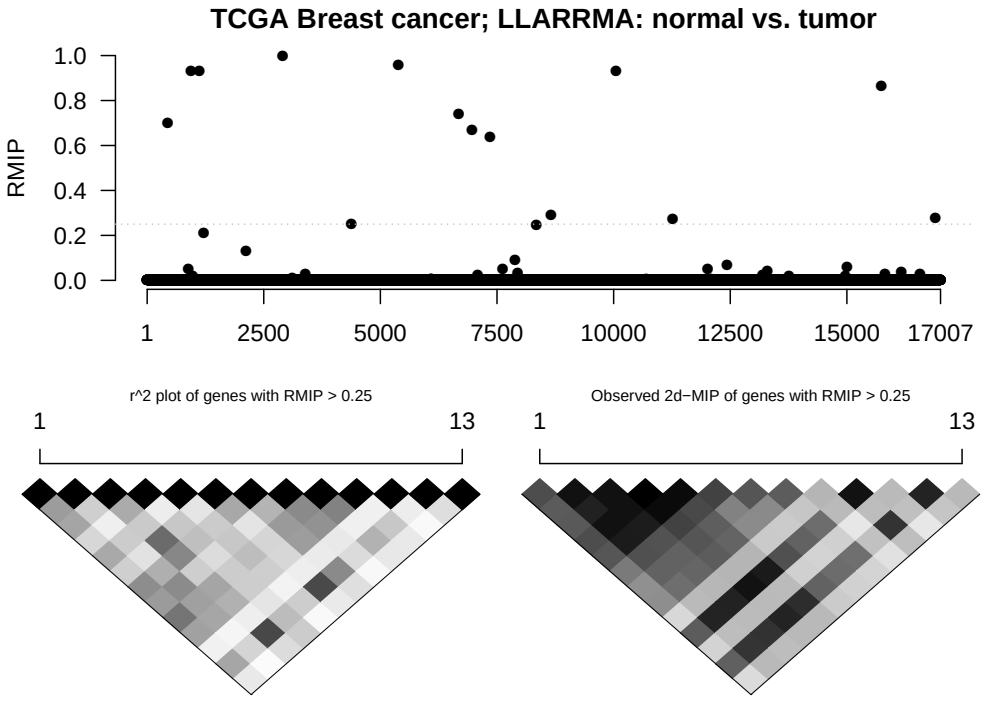

Figure 6.2: RMA based analyses of TCGA breast cancer. (Top) displays the LLARRMA output. (Bottom right) displays the r^2 of variables with RMIPs above 0.25. (Bottom left) displays the 2D-RMIP for the same set of variables.

and LuminalB, to try to identify cancer genes which help differentiate between the two similar subtypes. Figure 6.3 displays (top) the LLARRMA output with a gray dashed line indicating the RMIP threshold for 2D-RMIP comparisons and (bottom left) the r^2 values and (bottom right) 2D-RMIPs for variables with RMIPS greater than 0.25. When comparing the correlation of the GE values we observe that most of the highlighted variables from LLARRMA, those with RMIPs above 0.25, the majority of the pairwise gene correlations are quite low and do not seem to be highly related. When we examine the 2D-RMIP plot, we observe that many of the genes appear to be related with respect to the luminal subtype phenotype.

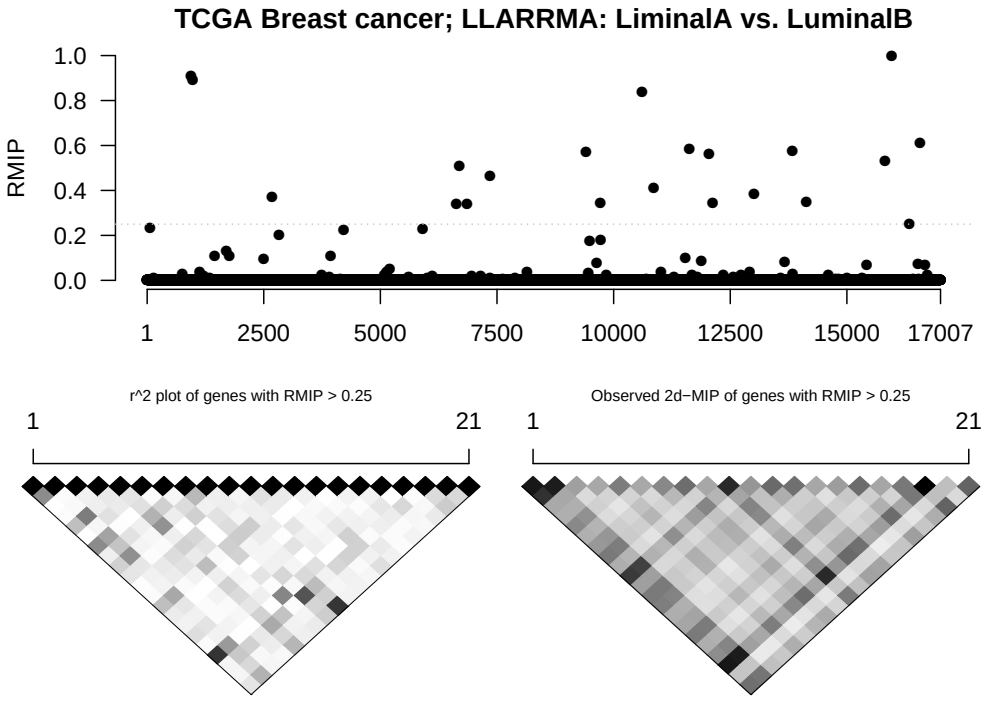

Figure 6.3: RMA based analyses of TCGA breast cancer luminal subtypes. (Top) displays the LLARRMA output. (Bottom right) displays the r^2 of variables with RMIPs above 0.25. (Bottom left) displays the 2D-RMIP for the same set of variables.

6.2 Discussion

In this chapter we presented the 2D-RMIP, a higher dimension summary of the information obtained from a RMA procedure, which can be used to show the relationships between predictors with respect to the response. The pairwise correlation between predictors contains useful information about the data set, but fails to capture information which is specific to a response which you are analyzing. The 2D-RMIP can be viewed as a measure of the relatedness of two variables with respect to the response of interest. This additional information comes out of the RMA procedure with little additional computation after filtering for variables which show evidence of relevance based on the RMIP.

We demonstrate one useful aspect of the 2D-RMIP, response specific relationships, on the TCGA breast cancer data. When we examine either healthy versus cancer

tissues, or LuminalA versus LuminalB cancer subtypes, that the 2D-RMIP displays relationships between variables relevant to the response that are not observable solely by the correlation structure. The 2D-RMIP also is able to show how some variables which are related based on correlation, are not related based on the model for the response.

The response specific variable relationships from the 2D-RMIPs are one useful extension that comes out of the RMA procedure. We believe that the 2D-RMIP may have further uses. The 2D-RMIP may also be useful in detecting ambiguities within the variable selection procedure over subsamples. Specifically, we mean that we may be able to detect pairs of variables that the variable selection procedure tends to include more or less often than we might expect based on the marginal RMIPs. One example of the type of confounding we may be able to detect with the 2D-RMIP would be variables that are nearly perfectly correlated to where the variable selection method essentially treats them as one variable and would tend to either include or exclude them both together. Another example of what we expect can be found from the 2D-RMIP is if the variable selection method is is switching between the inclusion of two related variables, including only one of the two on each subsample.

We have shown that the 2D-RMIP contains additional useful information which is not observable in the standard RMA output, the RMIP. We examined how the 2D-RMIP can be used to show relationships between variables relevant to the response of interest. We also discussed future extensions which can use the 2D-RMIP to detect problems within the variable selection procedure.

Chapter 7

Conclusions

The overarching theme of the work described in this dissertation is the use of LASSO-based resample model averaging to identify variables within a sparse complex model. In particular, we focus on datasets of complex responses that arise in genomics, while presenting the generality of the methodology which may be applied to a variety of disciplines. While there are many specific statistical methods that are designed for the analysis of such datasets, few beyond the simplistic single locus approaches are ever used in practice. The flexible methods described in this dissertation provide important improvements to identifying underling true signals in these datasets.

Generalized resample model averaging, the primary methodological contribution of this dissertation, provides a powerful new approach to analyzing association data. LLARRMA and its generalizations (LLARRMA-dawg and LLARRMA-haplo) provide a general framework which can be used to help better understand the underlying signals in complex models. We demonstrated how they can be applied to a wide variety of genetic association mappings, from human GWAS hit regions to genomewide association mapping in model organism crosses. We also have emphasized the generality of the approach, with the wide variety of models it may be applied to, which may allow for its uses beyond genetic association mapping.

Each of the projects presented here have strong potential for future research. We briefly summarize the primary directions of future research, described elsewhere in this dissertation, below:

- Investigate ways that we can incorporate prior information about variables associations into the LLARRMA framework.
- Investigate how to modify LLARRMA to analyze multiple independent data sets one the same regions simultaneously.
- Further investigate the model used in LLARRMA-haplo to improve the performance in some settings.
- Further investigate the usefulness of higher dimension summaries, such as the 2D-RMIP, of RMA procedures.
- Investigate alternative ways to improve the false positive bounds for generalized RMA.

The analysis of complex high-dimensional genomic data is a developing and very active field of research. There are many statistical challenges that are beyond the scope of this dissertation. Scientific questions in genomics and related fields that involve investigating the underlying model of the data are arising rapidly. It is important that statistical methodology not only keep pace with these problems, but are also general enough so that they can provide a basic framework for a wide variety of problems as they arise.

Chapter A

Appendix

A.1 Proofs for subsampling-based RMA Error Bound

The proof of the main theorems require the following lemmas.

Lemma 1. *For any set $K \subseteq \{1, \ldots, p\}$, a lower bound for the simultaneous selection probability is given by*

$$\hat{\Pi}_K^{\text{simult}} \geq 2\hat{\Pi}_K - 1.$$

Lemma 2. *Suppose $K \subseteq \{1, \ldots, p\}$. Let $\hat{S}$ be the the set of selected variables based on random weights $W = (w_1, \ldots, w_p) \sim$ Subsampling($\frac{1}{2}$). If $P(K \subseteq \hat{S}) \leq \epsilon$, then*

$$P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq \frac{\epsilon^2}{\xi}$$

Proof of Theorem 3.1. Define $\tilde{N} = N \cap \hat{S}$ to be the set of selected noise variables, and analogously $\tilde{S} = S \cap \hat{S}$. We will first show that for $k \in N$, $P(k \in \hat{S}) \leq \frac{q}{p}$. Write the expected number of falsely selected variables as $E(|\tilde{N}|) = E(|\hat{S}|) - E(|\tilde{S}|)$. Using the assumption that the selection procedure is no worse than random guessing, it follows that $E(|\tilde{S}|) \geq E(|\tilde{N}|)\frac{|S|}{|N|}$. Putting these together we have, $(1 + \frac{|S|}{|N|})E(|\tilde{N}|) \leq q$,

and hence $|N|^{-1}E(|\tilde{N}|) \leq \frac{q}{p}$. Using the exchangeability assumption, for all $k \in N$, $P(k \in \hat{S}) = \frac{E(|\tilde{N}|)}{|N|}$. Hence for $k \in N$, it holds that $P(k \in \hat{S}) \leq \frac{q}{p}$, as desired. (Note that this result is independent to the sample size used in the contribution of $\hat{S}$.) Using Lemma 2, it follows that for $k \in N$, $P(\hat{\Pi}_k^{\text{simult}} \geq \xi) \leq \frac{(\frac{q}{p})^2}{\xi}$ for $\xi \in (0,1)$. Applying Lemma 1, it follows that $P(\hat{\Pi}_k \geq \pi_{\text{thr}}) \leq P\left(\frac{\hat{\Pi}^{\text{simult}}+1}{2} \geq \pi_{\text{thr}}\right) \leq \frac{1}{2\pi_{\text{thr}}-1}(\frac{q}{p})^2$. Hence,

$$E(V) = \sum_{k \in N} P(\hat{\Pi}_k \geq \pi_{\text{thr}}) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q^2}{p}.$$

□

Proof of the Lemma 1. Let W_1, W_2 be complement random weights, where $E(W_i) = \frac{1}{2}\mathbf{1}$ for $i = 1, 2$. Denote by $s_K(\{1,1\})$ the probability $P^\star[\{K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1)\} \cap \{K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}]$, where the probability $P^\star$ is with respect to the random subsamples of half the data. The probabilities $s_K(\{1,0\})$, $s_K(\{0,1\})$, and $s_K(\{0,0\})$ are defined equivalently by $P^\star[\{K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1)\} \cap \{K \not\subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}]$, $P^\star[\{K \not\subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1)\} \cap \{K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}]$, and $P^\star[\{K \not\subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1)\} \cap \{K \not\subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}]$. Note that $\hat{\Pi}_K^{\text{simult}} = s_K(\{1,1\})$ and

$$\hat{\Pi}_K = s_K(\{1,0\}) + s_K(\{1,1\}) = s_K(\{0,1\}) + s_K(\{1,1\}),$$

$$1 - \hat{\Pi}_K = s_K(\{0,1\}) + s_K(\{0,0\}) = s_K(\{1,0\}) + s_K(\{0,0\}).$$

Under the assumption of a symmetric weighting distribution, it is obvious that $s_K(\{1,0\}) = s_K(\{0,1\})$. As $s_K(\{0,0\}) \geq 0$, it also follows that $s_K(\{1,0\}) \leq 1 - \hat{\Pi}_K$. Hence

$$\hat{\Pi}_K^{\text{simult}} = s_K(\{1,1\}) = \hat{\Pi}_K - s_K(\{1,0\}) \geq 2\hat{\Pi}_K - 1.$$

□

Proof of the Lemma 2. Let W_1, W_2 be complement random weights, where $\mathbf{W}_1 \sim \text{Subsampling}(\frac{1}{2})$. Denote the data, the n samples, by $\mathbf{Z}$. Define the binary random variable H_K for all subsets $K \subseteq \{1, \ldots, p\}$ as

$$H_K := \mathbb{I}_{\{K \subseteq \{\hat{S}(\mathbf{Z}, \mathbf{W}_1) \cap \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}\}}.$$

The simultaneous selection probability $\hat{\Pi}_K^{\text{simult}} = E^\star(H_K) = E(H_K|\mathbf{Z})$, where the expectation $E^\star$ is with respect to the random weighting of the samples and any randomization in the selection procedure of $\hat{S}$.

Using the conditional independence of $\hat{S}(\mathbf{Z}, \mathbf{W}_1)$ and $\hat{S}(\mathbf{Z}, \mathbf{W}_2)$ given the data $\mathbf{Z}$, the inequality $P(K \subseteq \hat{S}) \leq \epsilon$ implies that $P(H_K = 1) \leq P(K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1))^2 \leq \epsilon^2$. Therefore, $E(H_K) = E\{E(H_K|\mathbf{Z})\} = E(\hat{\Pi}_K^{\text{simult}}) \leq \epsilon^2$. Using Markov's inequality, we have that $\xi P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq E(\hat{\Pi}_K^{\text{simult}}) \leq \epsilon^2$. Thus, $P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq \frac{\epsilon^2}{\xi}$. □

Proof of Theorem 3.2. The proof follows similarly to that of Theorem 3.2, but rather than first showing that for $k \in N$, $P(k \in \hat{S}) \leq \frac{q}{p}$, we will use the new better than random guessing assumption to obtain a tighter bound at this step. Once we establish that for $k \in N$, $P(k \in \hat{S}) \leq \frac{q}{p+(\gamma-1)|S|}$, the proof follows the same steps. To establish this inequality, write the expected number of falsely selected variables as $E(|\tilde{N}|) = E(|\hat{S}|) - E(|\tilde{S}|)$. Using the selection procedure is γ better than random guessing, it follows that $E(|\tilde{S}|) \geq E(|\tilde{N}|)\gamma\frac{|S|}{|N|}$. Putting these together we have, $(1 + \gamma\frac{|S|}{|N|})E(|\tilde{N}|) \leq q$, and hence $|N|^{-1}E(|\tilde{N}|) \leq \frac{q}{p+(\gamma-1)|S|}$. The remainder follows the same steps as the original theorem. □

A.2 Proofs for generalized RMA Error Bound

Lemma 3. *Let $\hat{S}$ be the set of selected variables based on random weights $W = (w_1, \ldots, w_n)$ such that $E(w_i) = \frac{1}{2}$ for $i = 1, \ldots, n$. For any $k \in \{1, \ldots, p\}$, if $P(k \subseteq \hat{S}) \leq \epsilon$, then*

$$P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq \frac{\epsilon}{\xi}$$

Proof of the Lemma 3. The proof follows exactly as that of Lemma 2 until the inequality $P(K \subseteq \hat{S}) \leq \epsilon$ implies that $P(H_k = 1) \leq P(K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1))^2 \leq \epsilon^2$. Under the original setting we obtain this final inequality based on conditional independence of the selected sets, but here we are only able to obtain the inequality $P(H_k = 1) = P(K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_1) \cap K \subseteq \hat{S}(\mathbf{Z}, \mathbf{W}_2)) \leq \epsilon$ without making further assumptions. Therefore, $E(H_K) = E\{E(H_K|\mathbf{Z})\} = E(\hat{\Pi}_K^{\text{simult}}) \leq \epsilon$. Using Markov's inequality, we have that $\xi P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq E(\hat{\Pi}_K^{\text{simult}}) \leq \epsilon$. Thus, $P(\hat{\Pi}_K^{\text{simult}} \geq \xi) \leq \frac{\epsilon}{\xi}$. □

Proof of Theorem 3.3. The proof starts as in Theorem 3.2, yielding that for $k \in N$, it holds that $P(k \in \hat{S}) \leq \frac{q}{p+(\gamma-1)|S|}$. Rather than using Lemma 2 which assumed subsampling, we now use Lemma 3 where it follows that for $k \in N$, $P(\hat{\Pi}_k^{\text{simult}} \geq \xi) \leq \frac{(\frac{q}{p+(\gamma-1)|S|})}{\xi}$ for $\xi \in (0, 1)$. Using Lemma 1, it follows that $P(\hat{\Pi}_k \geq \pi_{\text{thr}}) \leq P\left(\frac{\hat{\Pi}^{\text{simult}}+1}{2} \geq \pi_{\text{thr}}\right) \leq \frac{1}{2\pi_{\text{thr}}-1}\frac{q}{p+(\gamma-1)|S|}$. Hence,

$$E(V) = \sum_{k \in N} P(\hat{\Pi}_k \geq \pi_{\text{thr}}) \leq \frac{1}{2\pi_{\text{thr}} - 1}\frac{pq}{p + (\gamma - 1)|S|}.$$

□

For the proof of Theorem 3.4, we need the following lemma. Lemma 3 is a single variable version of Lemma 2 under the added assumption of a monotone selection procedure.

Lemma 4. *Assume a monotone selection procedure with respect to weights. Let* $\hat{S}$ *be the set of selected variables based on random weights* $W = (w_1, \ldots, w_n)$ *such that* $E(w_i) = \frac{1}{2}$ *for* $i = 1, \ldots, n$. *For any* $k \in \{1, \ldots, p\}$, *if* $P(k \subseteq \hat{S}) \leq \epsilon$, *then*

$$P(\hat{\Pi}_k^{\text{simult}} \geq \xi) \leq \frac{\epsilon^2}{\xi}$$

Proof of the Lemma 4. The proof follows exactly as that of Lemma 2 with subsets K replaced by variables k until the inequality $P(k \in \hat{S}) \leq \epsilon$ implies that $P(H_k = 1) \leq P(k \in \hat{S}(\mathbf{Z}, \mathbf{W}_1))^2 \leq \epsilon^2$. Under the original setting we obtain this final inequality based on conditional independence of the selected sets, but here we are able to obtain the inequality based on the monotone inclusion condition. Note that $P(H_k = 1) = P(k \in \hat{S}(\mathbf{Z}, \mathbf{W}_1) \cap k \in \hat{S}(\mathbf{Z}, \mathbf{W}_2)) = E(E(\mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W}_1)\}} \mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W}_2)\}} | \mathbf{Z}))$. Defining $f(\mathbf{W}) = \mathbb{I}_{\{k \in \hat{S}(\mathbf{Z}, \mathbf{W})\}}$, we have $E(E(f(\mathbf{W}_1) f(\mathbf{W}_2)|Z)) = E(E(f(\mathbf{W}_1) f(1 - \mathbf{W}_1))|Z)$. Applying the monotone assumption, we have that if $f(\mathbf{W}_1)$ is monotonically non-increasing then $f(1 - \mathbf{W}_1)$ is monotonically non-decreasing and by a multivariate Chebyshev's covariance inequality we have $P(H_k = 1) \leq E(E(f(\mathbf{W}_1)|Z))^2 = P(k \in \hat{S}(\mathbf{Z}, \mathbf{W}_1))^2$ completing the part of the proof which differs under the general weighting function. □

Proof of Theorem 3.4. The proof follows exactly as Theorem 3.2 with the result from Lemma 2 replaced with the equivalent result from Lemma 4 under the additional assumption of a monotone selection procedure. □

Bibliography

Access Excellence. 2009. "Human Chromosomes." http://www.accessexcellence.org/RC/VL/GG/human.php.

Aitman, T J, A M Glazier, C A Wallace, L D Cooper, P J Norsworthy, F N Wahid, K M Al-Majali, P M Trembling, C J Mann, C C Shoulders and et al. 1999. "Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats." *Nature Genetics* 21(1):76–83.

Aitman, Timothy J, Charles Boone, Gary A Churchill, Michael O Hengartner, Trudy F C Mackay and Derek L Stemple. 2011. "The future of model organisms in human disease research." *Nature Reviews Genetics* 12(8):575–582.

Alexander, David H and Kenneth Lange. 2011. "Stability selection for genome-wide association." *Genetic Epidemiology* 35(7):722–8.

Amin, Najaf, Cornelia M van Duijn and Yurii S Aulchenko. 2007. "A genomic background based method for association analysis in related individuals." *PloS one* 2(12):e1274.

Ankeny, Rachel A. and Sabina Leonelli. 2011. "Whats so special about model organisms?" *Studies in History and Philosophy of Science Part A* 42(2):313 – 323.

Ansari, Anjum. 2001. "DNA structure." http://www.uic.edu/classes/phys/phys461/phys450/ANJUM04/.

Aranzana, M. J., S. Kim, K. Y. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. L. Tang, C. Toomajian, B. Traw, H. G. Zheng, J. Bergelson, C. Dean, P. Marjoram and M. Nordborg. 2005. "Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes." *Plos Genetics* 1(5):531–539.

Ayers, Kristin L and Heather J Cordell. 2010. "SNP Selection in genome-wide and candidate gene studies via penalized logistic regression." *Genetic epidemiology* 34(8):879–91.

Ayroles, Julien F, Mary Anna Carbone, Eric a Stone, Katherine W Jordan, Richard F Lyman, Michael M Magwire, Stephanie M Rollmann, Laura H Duncan, Faye Lawrence, Robert R H Anholt and Trudy F C Mackay. 2009. "Systems genetics of complex traits in Drosophila melanogaster." *Nature genetics* 41(3):299–307.

Balding, D.J. 2006. "A tutorial on statistical methods for population association studies." *Nature Reviews Genetics* 7(10):781–791.

Bennett, Brian J, Charles R Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus, Roumyana Yordanova, Bo Guan, Amy Truong, Wen-pin Yang, Aiqing He, Paul Kayne, Peter Gargalovic, Todd Kirchgessner, Calvin Pan, Lawrence W Castellani, Emrah Kostem, Nicholas Furlotte, Thomas A Drake, Eleazar Eskin and Aldons J Lusis. 2010. "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome research* 20(2):281–90.

Bink, Marco C a M and Fred a van Eeuwijk. 2009. "A Bayesian QTL linkage analysis of the common dataset from the 12th QTLMAS workshop." *BMC proceedings* 3 Suppl 1(Table 1):S4.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24:123–140.

Browning, Sharon R. 2008. "Missing data imputation and haplotype phase inference for genome-wide association studies." *Human Genetics* 124(5):439–450.

Buckland, ST, KP Burnham and NH Augustin. 1997. "Model selection: an integral part of inference." *Biometrics* 53(2):603–618.

Buckler, Edward S, James B Holland, Peter J Bradbury, Charlotte B Acharya, Patrick J Brown, Chris Browne, Elhan Ersoz, Sherry Flint-Garcia, Arturo Garcia, Jeffrey C Glaubitz, Major M Goodman, Carlos Harjes, Kate Guill, Dallas E Kroon, Sara Larsson, Nicholas K Lepak, Huihui Li, Sharon E Mitchell, Gael Pressoir, Jason a Peiffer, Marco Oropeza Rosas, Torbert R Rocheford, M Cinta Romay, Susan Romero, Stella Salvo, Hector Sanchez Villeda, H Sofia da Silva, Qi Sun, Feng Tian, Narasimham Upadyayula, Doreen Ware, Heather Yates, Jianming Yu, Zhiwu Zhang, Stephen Kresovich and Michael D McMullen. 2009. "The genetic architecture of maize flowering time." *Science (New York, N.Y.)* 325(5941):714–8.

Bühlmann, Peter and Bin Yu. 2002. "Analyzing bagging." *The Annals of Statistics* 30(4):927–961.

Calus, M P L and R F Veerkamp. 2007. "Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM." *Journal of animal breeding and genetics = Zeitschrift für Tierzüchtung und Züchtungsbiologie* 124(6):362–8.

Cantor, Rita M., Kenneth Lange and Janet S. Sinsheimer. 2010. "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application." *The American Journal of Human Genetics* 86(1):6 – 22.

Cavanagh, Colin, Matthew Morell, Ian Mackay and Wayne Powell. 2008. "From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants." *Current Opinion in Plant Biology* 11(2):215–221.

Cheng, Riyan, Mark Abney, Abraham A. Palmer and Andrew D. Skol. 2011. "QTLRel: an R Package for Genome-wide Association Studies in which Relatedness is a Concern." *Bmc Genetics* 12:66.

Cho, Seoae, Kyunga Kim, Young Jin Kim, Jong-Keuk Lee, Yoon Shin Cho, Jong-Young Lee, Bok-Ghee Han, Heebal Kim, Jurg Ott and Taesung Park. 2010. "Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis." *Annals of human genetics* 74(5):416–28.

Churchill, Gary a, David C Airey, Hooman Allayee, Joe M Angel, Alan D Attie, Jackson Beatty, William D Beavis, John K Belknap, Beth Bennett, Wade Berrettini, Andre Bleich, Molly Bogue, Karl W Broman, Kari J Buck, Ed Buckler, Margit Burmeister, Elissa J Chesler, James M Cheverud, Steven Clapcote, Melloni N Cook, Roger D Cox, John C Crabbe, Wim E Crusio, Ariel Darvasi, Christian F Deschepper, R W Doerge, Charles R Farber, Jiri Forejt, Daniel Gaile, Steven J Garlow, Hartmut Geiger, Howard Gershenfeld, Terry Gordon, Jing Gu, Weikuan Gu, Gerald de Haan, Nancy L Hayes, Craig Heller, Heinz Himmelbauer, Robert Hitzemann, Kent Hunter, Hui-Chen Hsu, Fuad a Iraqi, Boris Ivandic, Howard J Jacob, Ritsert C Jansen, Karl J Jepsen, Dabney K Johnson, Thomas E Johnson, Gerd Kempermann, Christina Kendziorski, Malak Kotb, R Frank Kooy, Bastien Llamas, Frank Lammert, Jean-Michel Lassalle, Pedro R Lowenstein, Lu Lu, Aldons Lusis, Kenneth F Manly, Ralph Marcucio, Doug Matthews, Juan F Medrano, Darla R Miller, Guy Mittleman, Beverly a Mock, Jeffrey S Mogil, Xavier Montagutelli, Grant Morahan, David G Morris, Richard Mott, Joseph H Nadeau, Hiroki Nagase, Richard S Nowakowski, Bruce F O'Hara, Alexander V Osadchuk, Grier P Page, Beverly Paigen, Kenneth Paigen, Abraham a Palmer, Huei-Ju Pan, Leena Peltonen-Palotie, Jeremy Peirce, Daniel Pomp, Michal Pravenec, Daniel R Prows, Zhonghua Qi, Roger H Reeves, John Roder, Glenn D Rosen, Eric E Schadt, Leonard C Schalkwyk, Ze'ev Seltzer, Kazuhiro Shimomura, Siming Shou, Mikko J Sillanpää, Linda D Siracusa, Hans-Willem Snoeck, Jimmy L Spearow, Karen Svenson, Lisa M Tarantino, David Threadgill, Linda a Toth, William Valdar, Fernando Pardo-Manuel de Villena, Craig Warden, Steve Whatley, Robert W Williams, Tim Wiltshire, Nengjun Yi, Dabao Zhang, Min Zhang and Fei Zou. 2004. "The Collaborative Cross, a community resource for the genetic analysis of complex traits." *Nature genetics* 36(11):1133–7.

Claeskens, Gerda and Fabrizio Consentino. 2008. "Variable selection with incomplete covariate data." *Biometrics* 64(4):1062–9.

Cleveland, Matthew a and Nader Deeb. 2009. "Evaluation of a genome-wide approach to multiple marker association considering different marker densities." *BMC proceedings* 3 Suppl 1:S5.

Cordell, Heather J and David G Clayton. 2002. "A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes." *Am J Hum Gen* 70(1):124–141.

Crooks, Lucy, Goutam Sahana, Dirk-Jan de Koning, Mogens SandøLund and Orjan Carlborg. 2009. "Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping." *BMC proceedings* 3 Suppl 1:S2.

Crow, James F. 2007. "Anecdotal, historical and critical commentaries on genetics. Gisela Mosig." *Genetics* 176:792–732.

Cubillos, Francisco a, Eleonora Billi, Enikö Zörgö, Leopold Parts, Patrick Fargier, Stig Omholt, Anders Blomberg, Jonas Warringer, Edward J Louis and Gianni Liti. 2011. "Assessing the complex architecture of polygenic traits in diverged yeast populations." *Molecular ecology* 20(7):1401–13.

Cule, Erika, Paolo Vineis and Maria De Iorio. 2011. "Significance testing in ridge regression for genetic data." *BMC Bioinformatics* 12(1):372+.

Darvasi, A. and M. Soller. 1995. "Advanced Intercross Lines, an Experimental Population for Fine Genetic-Mapping." *Genetics* 141(3):1199–1207.

Dastani, Zari, Marie-France Hivert, Nicholas Timpson, John R B Perry, Xin Yuan, Robert A Scott, Peter Henneman, Iris M Heid, Jorge R Kizer, Leo-Pekka Lyytikinen and et al. 2012. "Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals." *PLoS Genetics* 8(3):e1002607.

Demarest, K., J. McCaughran, E. Mahjubi, L. Cipp and R. Hitzemann. 1999. "Identification of an acute ethanol response quantitative trait locus on mouse chromosome 2." *Journal of Neuroscience* 19(2):549–561.

Devlin, B. and K. Roeder. 1999. "Genomic control for association studies." *Biometrics* 55(4):997–1004.

Doerge, Rebecca W. 2002. "Mapping and analysis of quantitative trait loci in experimental populations." *Nature reviews. Genetics* 3(1):43–52.

Fan, Jianqing and Jinchi Lv. 2008. "Sure Independence Screening for Ultra-High Dimensional Feature Space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5):849–911.

Fawcett, Tom. 2006. "An introduction to ROC analysis." *Pattern recognition letters* 27(8):861–874.

Flint, Jonathan and Eleazar Eskin. 2012. "Genome-wide association studies in mice." *Nature Reviews Genetics* 13(11):807–817.

Friedman, Jerome, Trevor Hastie and Rob Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software* 33(1):1.

Grupe, Andrew, Soren Germer, Jonathan Usuka and Dee Aud. 2001. "In Silico Mapping of Complex Disease-Related Traits in Mice." *Science* 292(5523):1915–1918.

Guan, Yongtao and Matthew Stephens. 2011. "Bayesian variable selection regression for genome-wide association studies and other large-scale problems." *Annals of Applied Statistics* 5(3):1780–1815.

Guy, Richard T, Peter Santago and Carl D Langefeld. 2012. "Bootstrap Aggregating of Alternating Decision Trees to Detect Sets of SNPs That Associate With Disease." *Library* 106(2):99–106.

Habier, D, R L Fernando and J C M Dekkers. 2007. "The impact of genetic relationship information on genome-assisted breeding values." *Genetics* 177(4):2389–97.

Haley, C S and S a Knott. 1992. "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers." *Heredity* 69(4):315–24.

Hansen, Bruce E. 2007. "Least Squares Model Averaging." *Econometrica* 75(4):1175–1189.

He, Qianchuan and Dan-Yu Lin. 2011. "A variable selection method for genome-wide association studies." *Bioinformatics* 27(1):1–8.

Hjort, Nils Lid and Gerda Claeskens. 2003. "Frequentist Model Average Estimators." *Journal of the American Statistical Association* 98(464):879–899.

Hoggart, CJ, JC Whittaker, Maria De Iorio and DJ Balding. 2008. "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies." *PLoS genetics* 4(7).

Howie, Bryan N, Peter Donnelly and Jonathan Marchini. 2009. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLoS Genetics* 5(6):e1000529.

Hu, Ying, Gang Wu, Michael Rusch, Luanne Lukes, Kenneth H Buetow, Jinghui Zhang and Kent H Hunter. 2012. "Integrated cross-species transcriptional network analysis of metastatic susceptibility." *Proceedings of the National Academy of Sciences of the United States of America* 109(8):3184–3189.

Jannink, J. L., M. C. A. M. Bink and R. C. Jansen. 2001. "Using complex plant pedigrees to map valuable genes." *Trends in plant science* 6(8):337–342.

Kang, Hyun Min, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti and Eleazar Eskin. 2010. "Variance component model to account for sample structure in genome-wide association studies." *Nature Genetics* 42(4):348–354.

Kang, Hyun Min, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly and Eleazar Eskin. 2008. "Efficient control of population structure in model organism association mapping." *Genetics* 178(3):1709–1723.

Kärkkinen, Hanni P and Mikko J Sillanpää. 2012. "Robustness of Bayesian multilocus association models to cryptic relatedness." *Annals of human genetics* 76(6):510–23.

Keating, Brendan J, Sam Tischfield, Sarah S Murray, Tushar Bhangale, Thomas S Price, Joseph T Glessner, Luana Galver, Jeffrey C Barrett, Struan F a Grant, Deborah N Farlow, Hareesh R Chandrupatla, Mark Hansen, Saad Ajmal, George J Papanicolaou, Yiran Guo, Mingyao Li, Stephanie Derohannessian, Paul I W de Bakker, Swneke D Bailey, Alexandre Montpetit, Andrew C Edmondson, Kent Taylor, Xiaowu Gai, Susanna S Wang, Myriam Fornage, Tamim Shaikh, Leif Groop, Michael Boehnke, Alistair S Hall, Andrew T Hattersley, Edward Frackelton, Nick Patterson, Charleston W K Chiang, Cecelia E Kim, Richard R Fabsitz, Willem Ouwehand, Alkes L Price, Patricia Munroe, Mark Caulfield, Thomas Drake, Eric Boerwinkle, David Reich, a Stephen Whitehead, Thomas P Cappola, Nilesh J Samani, a Jake Lusis, Eric Schadt, James G Wilson, Wolfgang Koenig, Mark I McCarthy, Sekar Kathiresan, Stacey B Gabriel, Hakon Hakonarson, Sonia S Anand, Muredach Reilly, James C Engert, Deborah a Nickerson, Daniel J Rader, Joel N Hirschhorn and Garret a Fitzgerald. 2008. "Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies." *PloS one* 3(10):e3583.

Kennedy, B. W., M. Quinton and J. A. M. Vanarendonk. 1992. "Estimation of Effects of Single Genes on Quantitative Traits." *Journal of animal science* 70(7):2000–2012.

Kim, Sulgi, Nathan J Morris, Sungho Won and Robert C Elston. 2010. "Single-marker and two-marker association tests for unphased case-control genotype data, with a power comparison." *Genetic Epidemiology* 34(1):67–77.

Kover, Paula X, William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M Ehrenreich, Michael D Purugganan, Caroline Durrant and Richard Mott. 2009. "A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana." *PLoS genetics* 5(7):e1000551.

Krzanowski, Wojtek J and David J Hand. 2009. *ROC Curves for Continuous Data*. 1st ed. Chapman & Hall/CRC.

Law, CN. 1966. "The location of genetic factors affecting aquantitative character in wheat." *Genetics* (March):487–498.

Ledur, Mônica Corrêa, Nicolas Navarro and Miguel Pérez-Enciso. 2009. "Data modeling as a main source of discrepancies in single and multiple marker association methods." *BMC proceedings* 3 Suppl 1:S9.

Lee, Sang Hong, Julius H. J. van der Werf, Ben J. Hayes, Michael E. Goddard and Peter M. Visscher. 2008. "Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data." *PLoS Genetics* 4(10):e1000231.

Li, Q., G. Zheng, X. Liang and K. Yu. 2009. "Robust Tests for Single-marker Analysis in Case-Control Genetic Association Studies." *Annals of Human Genetics* 73(2):245–252.

Li, Yun, Cristen J Willer, Jun Ding, Paul Scheet and Gonçalo R Abecasis. 2010. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genetic Epidemiology* 34(8):816–34.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics Wiley.

Malo, Nathalie, Ondrej Libiger and Nicholas J. Schork. 2008. "Accommodating Linkage Disequilibrium in Genetic-Association Analyses via Ridge Regression." *The American Journal of Human Genetics* 82(2):375–385.

Malosetti, M., C. G. van der Linden, B. Vosman and F. A. van Eeuwijk. 2007. "A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato." *Genetics* 175(2):879–889.

Manolio, TA, FS Collins, NJ Cox and DB Goldstein. 2009. "Finding the missing heritability of complex diseases." *Nature* 461(7265):747–753.

McClurg, Phillip, Jeff Janes, Chunlei Wu, David L Delano, John R Walker, Serge Batalov, Joseph S Takahashi, Kazuhiro Shimomura, Akira Kohsaka, Joseph Bass, Tim Wiltshire and Andrew I Su. 2007. "Genomewide association analysis in diverse inbred mice: power and population structure." *Genetics* 176(1):675–83.

McClurg, Phillip, Mathew T Pletcher, Tim Wiltshire and Andrew I Su. 2006. "Comparative analysis of haplotype association mapping algorithms." *BMC bioinformatics* 7:61.

Meier, Lukas. 2009. *grplasso: Fitting user specified models with Group Lasso penalty*. R package version 0.4-2.

Meier, Lukas, Sara Van De Geer and Peter Bühlmann. 2008. "The group lasso for logistic regression." *Journal of the Royal Statistical Society. Series B* 70(1):53–71.

Meinshausen, Nicolai and Peter Bühlmann. 2010. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4):417–473.

Mervis, Carolyn B, Joana Dida, Emily Lam, Nicole a Crawford-Zelli, Edwin J Young, Danielle R Henderson, Tuncer Onay, Colleen a Morris, Janet Woodruff-Borden, John Yeomans and Lucy R Osborne. 2012. "Duplication of GTF2I results in separation anxiety in mice and humans." *American journal of human genetics* 90(6):1064–70.

Morgan, Thomas Hunt, A. H. Sturtevant, H. J. Muller and Bridges C. B. 1915. *The Mechanism of Mendelian Heredity*. New York: Holt.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins and J. Flint. 2000. "A method for fine mapping quantitative trait loci in outbred animal stocks." *Proceedings of the National Academy of Sciences of the United States of America* 97(23):12649–12654.

Motyer, Allan J, Chris McKendry, Sally Galbraith and Susan R Wilson. 2011. "LASSO model selection with post-processing for a genome-wide association study data set." *BMC proceedings* 5 Suppl 9(Suppl 9):S24.

Müller, Bruno and Ueli Grossniklaus. 2010. "Model organisms–A historical perspective." *Journal of proteomics* 73(11):2054–63.

Neale, B.M., M Ferreira, S Medland and D Posthuma. 2008. *Statistical genetics: gene mapping through linkage and association*. Taylor & Francis Group.

Newton, Michael A. and Adrian E. Raftery. 1994. "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society. Series B (Methodological)* 56(1).

O'Hara, R. B. and M. J. Sillanpää. 2009. "A review of Bayesian variable selection methods: what, how and which." *Bayesian Analysis* 4(1):85–117.

Palmer, Abraham a and Harriet de Wit. 2011. "Translational genetic approaches to substance use disorders: bridging the gap between mice and humans." *Human genetics* .

Patterson, Nick, Alkes L. Price and David Reich. 2006. "Population structure and eigenanalysis." *Plos Genetics* 2(12):2074–2093.

Pikkuhookana, P and M J Sillanpää. 2009. "Correcting for relatedness in Bayesian models for genomic data association analysis." *Heredity* 103(3):223–37.

Pletcher, Mathew T, Philip McClurg, Serge Batalov, Andrew I Su, S Whitney Barnes, Erica Lagler, Ron Korstanje, Xiaosong Wang, Deborah Nusskern, Molly A Bogue,

Richard J Mural, Beverly Paigen and Tim Wiltshire. 2004. "Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse." *PLoS biology* 2(12):e393.

Politis, D.N., J.P. Romano and M. Wolf. 1999. *Subsampling*. Springer series in statistics Springer.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick and David Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* 38(8):904–909.

Price, Alkes L., Noah A. Zaitlen, David Reich and Nick Patterson. 2010. "New approaches to population stratification in genome-wide association studies." *Nature Reviews Genetics* 11(7):459–463.

Pritchard, J. K., M. Stephens and P. Donnelly. 2000. "Inference of population structure using multilocus genotype data." *Genetics* 155(2):945–959.

Pruim, Randall J, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonçalo R Abecasis and Cristen J Willer. 2010. "LocusZoom: regional visualization of genome-wide association scan results." *Bioinformatics (Oxford, England)* 26(18):2336–7.

Purcell, S, B Neale, K Toddbrown, L Thomas, M Ferreira, D Bender, J Maller, P Sklar, P Debakker and M Daly. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81(3):559–575.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *Annals of Statistics* 9(1):130–134.

Scheet, Paul and Matthew Stephens. 2006. "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." *The American Journal of Human Genetics* 78(4):629–644.

Schomaker, Michael, Alan T.K. Wan and Christian Heumann. 2010. "Frequentist Model Averaging with missing observations." *Computational Statistics & Data Analysis* 54(12):3336–3347.

Schön, Chris C, H Friedrich Utz, Susanne Groh, Bernd Truberg, Steve Openshaw and Albrecht E Melchinger. 2004. "Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits." *Genetics* 167(1):485–98.

Servin, Bertrand and Matthew Stephens. 2007. "Imputation-based analysis of association studies: candidate regions and quantitative traits." *PLoS Genetics* 3(7):1296–1308.

Shah, Rajen and Richard J Samworth. 2011. Variable selection with error control: Another look at Stability Selection. Technical Report arXiv:1105.5578.

Shi, G, E Boerwinkle and AC Morrison. 2011. "Mining gold dust under the genome wide significance level: a twostage approach to analysis of GWAS." *Genetic Epidemiology* 35(2):111–118.

Sillanpää, M J. 2011. "Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses." *Heredity* 106(4):511–9.

Sillanpää, Mikko J and Madhuchhanda Bhattacharjee. 2005. "Bayesian association-based fine mapping in small chromosomal segments." *Genetics* 169(1):427–39.

Siva, Nayanah. 2008. "1000 Genomes project." *Nature Biotechnology* 26(3):256.

Solberg, Leah C, William Valdar, Dominique Gauguier, Graciela Nunez, Amy Taylor, Stephanie Burnett, Carmen Arboledas-Hita, Polinka Hernandez-Pliego, Stuart Davidson, Peter Burns, Shoumo Bhattacharya, Tertius Hough, Douglas Higgs, Paul Klenerman, William O Cookson, Youming Zhang, Robert M Deacon, J Nicholas P Rawlins, Richard Mott and Jonathan Flint. 2006. "A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice." *Mammalian genome : official journal of the International Mammalian Genome Society* 17(2):129–46.

Stephens, Matthew and David J Balding. 2009. "Bayesian statistical methods for genetic association studies." *Nature reviews. Genetics* 10(10):681–90.

Strange, Amy, Francesca Capon, Chris C A Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Céline Bellenguez, Judith G M Bergboer, Jenefer M Blackwell, Elvira Bramon, Suzannah J Bumpstead, Juan P Casas, Michael J Cork, Aiden Corvin, Panos Deloukas, Alexander Dilthey, Audrey Duncanson, Sarah Edkins, Xavier Estivill, Oliver Fitzgerald, Colin Freeman, Emiliano Giardina, Emma Gray, Angelika Hofer, Ulrike Hüffmeier, Sarah E Hunt, Alan D Irvine, Janusz Jankowski, Brian Kirby, Cordelia Langford, Jesús Lascorz, Joyce Leman, Stephen Leslie, Lotus Mallbris, Hugh S Markus, Christopher G Mathew, W H Irwin McLean, Ross McManus, Rotraut Mössner, Loukas Moutsianas, Asa T Naluai, Frank O Nestle, Giuseppe Novelli, Alexandros Onoufriadis, Colin N a Palmer, Carlo Perricone, Matti Pirinen, Robert Plomin, Simon C Potter, Ramon M Pujol, Anna Rautanen, Eva Riveira-Munoz, Anthony W Ryan, Wolfgang Salmhofer, Lena Samuelsson, Stephen J Sawcer, Joost Schalkwijk, Catherine H Smith, Mona Stå hle, Zhan Su, Rachid Tazi-Ahnini, Heiko Traupe, Ananth C Viswanathan, Richard B

Warren, Wolfgang Weger, Katarina Wolk, Nicholas Wood, Jane Worthington, Helen S Young, Patrick L J M Zeeuwen, Adrian Hayday, a David Burden, Christopher E M Griffiths, Juha Kere, André Reis, Gilean McVean, David M Evans, Matthew A Brown, Jonathan N Barker, Leena Peltonen, Peter Donnelly and Richard C Trembath. 2010. "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." *Nature Genetics* 42(11):985–990.

Su, Zhan, Jonathan Marchini and Peter Donnelly. 2011. "HAPGEN2: simulation of multiple disease SNPs." *Bioinformatics* 27(16):1–2.

Svenson, Karen L, Daniel M Gatti, William Valdar, Catherine E Welsh, Riyan Cheng, Elissa J Chesler, Abraham A Palmer, Leonard McMillan and Gary A Churchill. 2012. "High-resolution genetic mapping using the Mouse Diversity outbred population." *Genetics* 190(2):437–47.

Tanaka, Toshihiro. 2009. "HapMap project." *Nippon Rinsho* 67(6):1068–1071.

TCGA. 2012. "Comprehensive molecular portraits of human breast tumours." *Nature* 490(7418):61–70.

Threadgill, David W. 2006. "Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics." *Mammalian genome : official journal of the International Mammalian Genome Society* 17(1):2–4.

Threadgill, David W, Kent W Hunter and Robert W Williams. 2002. "Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort." *Mammalian genome : official journal of the International Mammalian Genome Society* 13(4):175–8.

Tibshirani, R. 1996. "Regression shrinkage and selection via lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Utz, Hf, Ae Melchinger and Cc Schön. 2000. "Bias and Sampling Error of the Estimated Proportion of Genotypic Variance Explained by Quantitative Trait Loci Determined From Experimental Data in Maize Using Cross Validation and Validation With Independent Samples." *Genetics* 154(3):1839–1849.

Valdar, William, Christopher C Holmes, Richard Mott and Jonathan Flint. 2009. "Mapping in structured populations by resample model averaging." *Genetics* 182(4):1263–77.

Valdar, William, Jeremy Sabourin, Andrew Nobel and Christopher C Holmes. 2012. "Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging." *Genetic epidemiology* 36(5):451–62.

Valdar, William, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott and Jonathan Flint. 2006. "Genome-wide genetic association of complex traits in heterogeneous stock mice." *Nature genetics* 38(8):879–87.

Veenstra-VanderWeele, Jeremy, Asfia Qaadir, Abraham a Palmer, Edwin H Cook and Harriet de Wit. 2006. "Association between the casein kinase 1 epsilon gene region and subjective response to D-amphetamine." *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 31(5):1056–63.

Wang, Xuefeng, Nathan J Morris, Daniel J Schaid and Robert C Elston. 2012. "Power of single- vs. multi-marker tests of association." *Genetic Epidemiology* 36(5):480–7.

Warren, Liling L., Li Li, Matthew R. Nelson, Margaret G. Ehm, Judong Shen, Dana J. Fraser, Jennifer L. Aponte, Keith L. Nangle, Andrew J. Slater, Peter M. Woollard, Matt D. Hall, Simon D. Topp, Xin Yuan, Lon R. Cardon, Stephanie L. Chissoe, Vincent Mooser, Andrew D. Morris, Colin N.A. Palmer, John R. Perry, Timothy M. Frayling, John C. Whittaker and Dawn M. Waterworth. 2012. "Deep Resequencing Unveils Genetic Architecture of ADIPOQ and Identifies a Novel Low-Frequency Variant Strongly Associated With Adiponectin Variation." *Diabetes* 61(5):1297–1301.

Waterston, Robert H, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An and et al. 2002. "Initial sequencing and comparative analysis of the mouse genome." *Nature* 420(6915):520–562.

Williams, Christopher J and Joe C Christian. 2006. "Frequentist model-averaged estimators and tests for univariate twin models." *Behavior genetics* 36(5):687–96.

Wiltshire, Tim, Mathew T. Pletcher, Serge Batalov, S. Whitney Barnes, Lisa M. Tarantino, Michael P. Cooke, Hua Wu, Kevin Smylie, Andrey Santrosyan, Neal G. Copeland, Nancy A. Jenkins, Francis Kalush, Richard J. Mural, Richard J. Glynne, Steve A. Kay, Mark D. Adams and Colin F. Fletcher. 2003. "Genome-Wide Single-Nucleotide Polymorphism Analysis Defines Haplotype Patterns in Mouse." *Proceedings of the ...* 100(6):3380–3385.

WTCCC. 2007. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447(7145):661–78.

Wu, T.T., Y.F. Chen, Trevor Hastie, Eric Sobel and Kenneth Lange. 2009. "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics* 25(6):714.

Yalcin, B., J. Fullerton, S. Miller, D. A. Keays, S. Brady, A. Bhomra, A. Jefferson, E. Volpi, R. R. Copley, J. Flint, R. Mott and David Housman. 2004. "Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice." *Proceedings of the ...* 101(24):9734–9739.

Yeager, Meredith, Nick Orr, Richard B. Hayes, Kevin B. Jacobs, Peter Kraft, Sholom Wacholder, Mark J. Minichiello, Paul Fearnhead, Kai Yu, Nilanjan Chatterjee, Zhaoming Wang, Robert Welch, Brian J. Staats, Eugenia E. Calle, Heather S. Feigelson, Michael J. Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R. Schumacher, Edward Giovannucci, Walter C. Willett, Geraldine Cancel-Tassin, Olivier Cussenot, Antoine Valeri, Gerald L. Andriole, Edward P. Gelmann, Margaret Tucker, Daniela S. Gerhard, Joseph F. Fraumeni, Robert Hoover, David J. Hunter, Stephen J. Chanock and Gilles Thomas. 2007. "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24." *Nature Genetics* 39(5):645–649.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich and Edward S Buckler. 2006. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nature genetics* 38(2):203–8.

Yuan, M. and Yi Lin. 2006*a*. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Yuan, Ming and Yi Lin. 2006*b*. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, Jianming Yu, Donna K. Arnett, Jose M. Ordovas and Edward S. Buckler. 2010. "Mixed linear model approach adapted for genome-wide association studies." *Nature genetics* 42(4):355–U118.

Zhao, Keyan, Maria Jose Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram and Magnus Nordborg. 2007. "An Arabidopsis example of association mapping in structured samples." *Plos Genetics* 3(1):e4.

Zheng, Jin, Yun Li, Gonçalo R. Abecasis and Paul Scheet. 2011. "A comparison of approaches to account for uncertainty in analysis of imputed genotypes." *Genetic Epidemiology* 35(2):102–110.

Zhou, H, D H Alexander, M E Sehl, J S Sinsheimer, E M Sobel and K Lange. 2011. "Penalized regression for genome-wide association screening of sequence data." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* pp. 106–17.

Zhou, Hua, Mary E Sehl, Janet S Sinsheimer and Kenneth Lange. 2010. "Association screening of common and rare genetic variants by penalized regression." *Bioinformatics (Oxford, England)* 26(19):2375–2382.

Zhou, Xiang and Matthew Stephens. 2012. "Genome-wide efficient mixed-model analysis for association studies." *Nature Genetics* 44(7):821–4.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101(476):1418–1429.

Zuber, Verena, APD Silva and Korbinian Strimmer. 2012. "A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies." *BMC Bioinformatics* 13(284):2–9.